

AD-A110 535

RAND CORP SANTA MONICA CA  
GENERALIZABILITY THEORY: 1973-1980, (U)  
JUL 81 R J SHAVELSON, N M WEBB

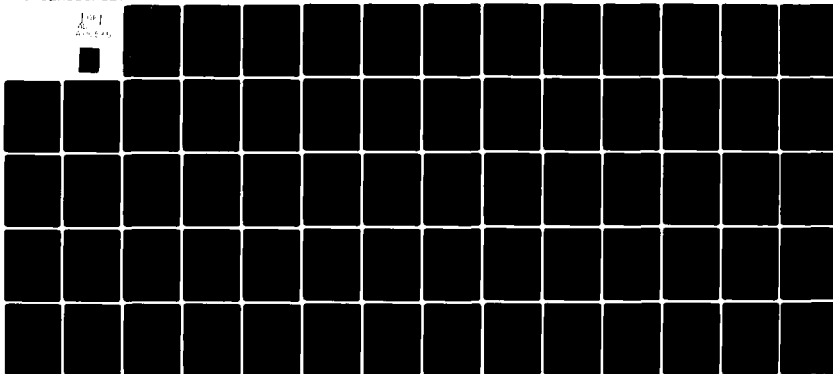
F/G 12/1

UNCLASSIFIED

RAND/P-6560

NL

TOP  
SECRET



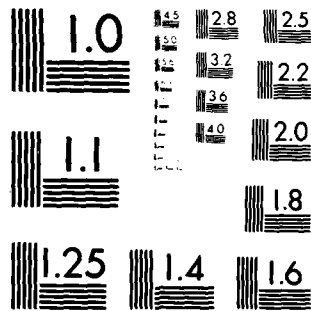
END

DATE

FILED



DTIC



MICROCOPY RESOLUTION TEST CHART  
NATIONAL BUREAU OF STANDARDS 1963-A

AD/P-6580

GENERALIZABILITY THEORY, 1973-1980

LEVEL II

(2)

Yw

AD A110535

GENERALIZABILITY THEORY: 1973-1980

Richard J. Shavelson  
Noreen M. Webb

DTIC  
COLLECTED  
FEB 5 1982  
H

July 1981

DTIC FILE COPY

DISTRIBUTION STATEMENT A  
Approved for public release;  
Distribution Unlimited

P-6580

82 02 04 1988

Test Results

Test results show that the system is capable of operating at a maximum speed of 1000 RPM. The test results also show that the system is capable of operating at a maximum torque of 1000 lb-ft. The test results also show that the system is capable of operating at a maximum power of 1000 hp.

Test results show that the system is capable of operating at a maximum speed of 1000 RPM. The test results also show that the system is capable of operating at a maximum torque of 1000 lb-ft. The test results also show that the system is capable of operating at a maximum power of 1000 hp.

1

GENERALIZABILITY THEORY:

1973-1980

Richard J. Shavelson  
University of California, Los Angeles  
and  
The Rand Corporation

and

Noreen M. Webb  
University of California, Los Angeles


British Journal of Mathematical and Statistical Psychology (in press)

DT  
ELEC  
1234  
E

PREFACE

In this paper we review generalizability (G) theory, a theory of the multifaceted errors of a behavioral measurement. The review was undertaken at the request of Philip Levy, then editor of the Journal. His idea was that the review would commemorate the first article on G theory, which the Journal published in 1963 (Cronbach, Rajaratnam, & Gleser, 1963). For these and personal reasons, we undertook the review. The review does not cover the period 1963-1972 because that has already been done by Cronbach, Gleser, Nanda, and Rajaratnam (1972).

In Section 1 we sketch out generalizability theory for those who are not familiar with it. In doing so, we summarize the notation used in the review. Section 2 reviews theoretical contributions. While it primarily reflects what has been published, we take up some new topics and identify others in need of treatment. Section 3 presents an application of the theory in some detail. This application illustrates basic concepts in the theory (Section 1) as well as recent theoretical contributions (Section 2).



Accession For	
NTIS	<input checked="" type="checkbox"/>
DTIC	<input type="checkbox"/>
Unannounced	<input type="checkbox"/>
Continuation	
By	
Distribution	
Availability Codes	
Dist	
A	

ABSTRACT

This paper reviews the developments in generalizability theory from 1973 to 1980. The first section presents a sketch of generalizability theory. The second section reviews theoretical contributions about (1) problems associated with estimating variance components, including sampling variability and negative estimates, (2) fixed facets, (3) criterion-referenced measurement, (4) symmetry, (5) multivariate generalizability, and (6) sampling in observational measurement. The final section presents an illustrative application of generalizability theory, including univariate and multivariate generalizability analyses of balanced and unbalanced designs, and Bayesian estimation of variance components.

# 1. SKETCH OF GENERALIZABILITY THEORY

Generalizability theory evolved out of the recognition that the concept of undifferentiated error in classical test theory provided too gross a characterization of the multiple sources of error in a measurement. The multidimensional nature of measurement error can be seen in how a test score is obtained. For example, one of many possible test forms might be administered in one of many possible occasions by one of many possible testers. Each of these choices--test form, occasion, and tester--is a potential source of error. G theory assesses each source of error in order to characterize the measurement and improve its design.

A behavioral measurement, then, is a sample from a universe of admissible observations, characterized by one or more facets (e.g., test forms, occasions, testers).<sup>1</sup> This universe is usually defined by the Cartesian product of the levels (called conditions in G theory) of the facets. From this perspective, Cronbach et al. (1972, p. 15) say:

The score on which the decision is to be based is only one of many scores that might serve the same purpose. The decision maker is almost never interested in the response given to the particular stimulus objects or questions, to the particular tester, at the particular moment of testing. Some, at least, of these conditions of measurement could be altered without making the score any less acceptable to the decision maker. That is to say, there is a universe of observations, any of which would have yielded a usable basis for the decision. The ideal datum on which to base the decision would be something like the person's mean score over all acceptable observations, which we shall call his "universe score." The investigator uses the observed score or some function of it as if it were the universe score. That is, he generalizes from sample to universe. The question of "reliability" thus resolves into a question of accuracy of generalization, or generalizability.

Since different measurements may represent different universes, G theory speaks of universe scores rather than true scores, acknowledging that there are different universes to which decisionmakers may generalize. Likewise, the theory speaks of generalizability coefficients rather than the reliability coefficient, realizing that the value of the coefficient may change as definitions of universes change.



In G theory, a person's score is decomposed into a component for the universe score ( $\mu_p$ ) and one or more error components. To illustrate this decomposition, we consider the simplest case for pedagogical purposes--a one facet,  $p \times i$  (person by, say, test form) design. (The object of measurement, here persons, is not a source of error and, therefore, is not a facet.) The presentation readily generalizes to more complex designs. In the  $p \times i$  design with generalization over all admissible test forms taken from an indefinitely large universe, the score for a particular person ( $p$ ) on a particular form ( $i$ ) is:

$$\begin{aligned}
 [1] \quad X_{pi} &= \mu && \text{(grand mean)} \\
 &+ \mu_p - \mu && \text{(person effect)} \\
 &+ \mu_i - \mu && \text{(form effect)} \\
 &+ X_{pi} - \mu_p - \mu_i + \mu. && \text{(residual)}
 \end{aligned}$$

Except for the grand mean, each score component has a distribution. Considering all persons in the population, there is a distribution of  $\mu_p - \mu$  with mean zero and variance  $\xi(\mu_p - \mu)^2 = \sigma_p^2$ , which is called the universe-score variance and is analogous to the true-score variance of classical theory. Similarly, the component for test form has mean zero and variance  $\xi(\mu_i - \mu)^2 = \sigma_i^2$  which indicates the variance of constant errors associated with test forms, while the residual component has mean zero and variance  $\sigma_{pi,e}^2$ , which indicates the person  $\times$  form interaction confounded with residual error, since there is one observation per cell. The collection of observed scores,  $X_{pi}$ , has a variance  $\sigma_{X_{pi}}^2 = \xi(X_{pi} - \mu)^2$  which equals the sum of the variance components:

$$[2] \quad \sigma_{X_{pi}}^2 = \sigma_p^2 + \sigma_i^2 + \sigma_{pi,e}^2$$

G theory focuses on these variance components. The relative magnitudes of the components provide information about particular sources of error influencing a measurement. It is convenient to estimate variance components from an ANOVA of sample data. Numerical estimates of the variance components are obtained by setting the expected mean squares equal to the observed mean squares and solving the set of simultaneous equations as shown in Table 1.

Table 1  
ESTIMATES OF VARIANCE COMPONENTS FOR A  
ONE FACET,  $p \times i$ , DESIGN

Source of Variation	Mean Square	Expected Mean Square *	Estimated Variance Component
Person (p)	$MS_p$	$\sigma_{pi,e}^2 + n_i \sigma_p^2$	$\hat{\sigma}_p^2 = (MS_p - MS_{res})/n_i$
Form (i)	$MS_i$	$\sigma_{pi,e}^2 + n_p \sigma_i^2$	$\hat{\sigma}_i^2 = (MS_i - MS_{res})/n_p$
$p \times i, e$	$MS_{res}$	$\sigma_{pi,e}^2$	$\hat{\sigma}_{pi,e}^2 = MS_{res}$

\*  $n_i$  = number of forms;  $n_p$  = number of persons.

Variance components are estimated by means of a generalizability (G) study. "The instrument developer, carrying out a G study to guide users of his instrument will, in the design of that study, treat systematically the facets that are likely to enter into generalizations of various users" (Cronbach et al., 1972, p. 21). Ordinarily, the universe of admissible observations is defined as broadly as possible within practical and theoretical constraints. In most cases, Cronbach et al. recommended using a crossed design so that all of the variance components can be estimated. Cronbach et al. (1972) noted, however, that a nested G study is sometimes useful because it provides more degrees of freedom for some estimates of variance components.

G theory distinguishes a decision (D) study from a G study. This distinction recognizes that certain studies are associated with the development of a measurement procedure (G studies) while other studies then apply the procedure (D studies). In planning the D study, the decisionmaker (a) defines the universe of generalization and (b) specifies his proposed interpretation of a measurement. These plans determine (c) the questions to be asked of the G study data in order to optimize the measurement design. Each of these points is considered in turn.

(a) G theory recognizes that the universe of admissible observations encompassed by a G study may be broader than the universe to which a decisionmaker wishes to generalize. That is, the decisionmaker proposes to generalize to a universe comprised of some subset of facets in the G study. This universe is called the universe of generalization. It may be defined by reducing the universe of admissible observations, i.e., by reducing the levels of a facet (creating a fixed facet; cf. fixed factor in ANOVA), by selecting and thereby controlling one level of a facet, or by ignoring a facet. All three alternatives have consequences for the estimation of the components of error variance that enter into the observed score variance.

(b) G theory recognizes that decision makers use the same test score in different ways. For example, some interpretations may focus on individual differences (i.e., relative or comparative decisions), some may use the observed score as an estimate of a person's universe score (absolute decisions; cf. criterion-referenced interpretations), while still others may use the observed score in a regression estimate of the universe score (cf. Kelley's, 1947, regression estimate of true scores). There is a different error associated with each of these proposed interpretations. For relative decisions, the error in a  $p \times i$  design is defined as:

$$[3] \quad \delta_{pI} = (X_{pI} - \mu_I) - (\mu_p - \mu),$$

where I indicates that an average has been taken over the levels of

facet  $i$  under which  $p$  was observed. The variance of the errors for relative decisions are:

$$[4] \quad \sigma_{\delta}^2 = \sigma_{pI}^2 = \sigma_{pi,e}^2 / n_i',$$

where  $n_i'$  indicates the number of conditions of facet  $i$  to be sampled in a D study. Notice that (a)  $\sigma_{pi,e}^2 / n_i'$  is the standard error of the mean of a person's scores averaged over the levels of  $i$  (test forms in our example). And (b) the magnitude of the error is under the control of the decisionmaker in the D study. In order to reduce  $\sigma_{\delta}^2$ ,  $n_i'$  may be increased. This is analogous to the Spearman-Brown prophecy formula in classical theory and the standard error of the mean in sampling theory.

For absolute decisions, the error is defined as:

$$[5] \quad \Delta_{pI} = X_{pI} - \mu_p.$$

The variance of these errors in a  $p \times i$  design is:

$$\sigma_{\Delta}^2 = \sigma_I^2 + \sigma_{pI}^2 = \sigma_i^2 / n_i + \sigma_{pi,e}^2 / n_i'.$$

In contrast to  $\sigma_{\delta}^2$ ,  $\sigma_{\Delta}^2$  includes the variance of constant errors associated with facet  $i$  ( $\sigma_i^2$ ). This arises because, in absolute decisions, the difficulty of the particular test forms that a person receives will influence his observed score and, hence, the decisionmaker's estimate of his universe score. For relative decisions, however, the effect of test form is constant for all persons and so does not influence the rank ordering of them (see Erlich & Shavelson, 1976b).

Finally, for decisions based on the regression estimate of a person's universe score, error (of estimate) is defined as:

$$[7] \quad \epsilon_p = \hat{\mu}_p - \mu_p,$$

where  $\hat{\mu}_p$  is the regression estimate of a person's universe score,  $\mu_p$ .

The estimation procedure for the variance of errors of estimate may be found in Cronbach et al. (1972, p. 97ff).

(c) D studies encompass a wide variety of designs including crossed, partially nested, and completely nested designs. All facets in the D design may be random (cf. random model) or only some may be random (cf. mixed model). Often, in D studies, nested designs are used for convenience, for increasing sample size, or both. Forms may be nested within persons (we write i:p to denote nesting). So, the effect of the constant errors associated with facet i is confounded with the effect associated with the person by i-facet interaction ( $\pi_{i,e}$ ). Hence.

$$[8] \quad \sigma_{X_{pI}}^2 = \sigma_p^2 + \sigma_{I,pI,e}^2 = \sigma_p^2 + \sigma_{\Delta}^2 .$$

Note that, for a completely nested design,  $\sigma_{\delta}^2 = \sigma_{\Delta}^2$ .

While stressing the importance of variance components and errors such as  $\sigma_{\delta}^2$ , G theory also provides a coefficient analogous to the reliability coefficient in classical theory. The generalizability coefficient,  $\xi_p^2$ , is defined as the ratio of the universe-score variance to the expected observed-score variance, i.e., an intraclass correlation:

$$[9] \quad \xi_p^2 = \frac{\sigma_p^2}{\xi \sigma^2(X)} = \frac{\sigma_p^2}{\sigma_p^2 + \sigma_{\delta}^2} .$$

The expected observed-score variance is used in G theory because the theory assumes only random sampling of the levels of facets and so the observed-score variance may change from one application of the design to another. Sample estimates of the parameters in [9] are used to estimate the G coefficient:

$$[9a] \quad \xi_p^2 = \frac{\hat{\sigma}_p^2}{\hat{\sigma}_p^2 + \hat{\sigma}_{\delta}^2} .$$

$\hat{\xi}_p^2$  is a biased but consistent estimator of  $\xi_p^2$ .

For absolute decisions a generalizability coefficient can be defined in an analogous manner:

$$[10] \quad \xi_p^2 = \frac{\sigma_p^2}{\sigma_p^2 + \sigma_\Delta^2}, \text{ and}$$

$$[10a] \quad \hat{\xi}_p^2 = \frac{\hat{\sigma}_p^2}{\hat{\sigma}_p^2 + \hat{\sigma}_\Delta^2}.$$

Finally, note that, for completely nested designs regardless of whether relative or absolute decisions are to be made, error variance is defined as  $\sigma_\Delta^2$  and so [10] provides the generalizability coefficient for such designs.

Generalizability theory has been applied widely in the behavioral sciences. It has been applied, for example, in studying the dependability of measures of the behavior of schizophrenic patients (e.g., Mariotto & Farrell, 1979), assertion in the elderly (Edinberg, Karoly, & Gleser, 1977), free-recall in children (Peng & Farr, 1976), depth and duration of sleep (Coates, Rosekind, Strossen, Thoresen, & Kirmil-Gray, 1979), behavior of teachers (e.g., Erlich & Shavelson, 1978), dentists' sensitivity toward patients (Gershen, 1976), educational attainment (Cardinet, Tourneur, & Allal, 1976a), job satisfaction using Spanish and English forms (Katerberg, Smith, & Hoy, 1977), student ratings of instruction (e.g., Gillmore, Kane, & Naccarato, 1978), and heterosexual social anxiety (Farrell, Marco, Conger, Curran, & Wallander, 1979).

A study of the dependability of measures of psychological improvement of disaster survivors (Gleser, Green, & Winget, 1978) illustrates the theory's treatment of multifaceted measurement error. Twenty adult survivors (S) were interviewed independently by two interviewers (I) in order to obtain data on the extent of psychiatric impairment resulting from the disaster. Two raters (R) "quantified" the interview data by rating each survivor on a number of subscales, such as anxiety, of the Psychiatric Evaluation Form. In differentiating survivors with

respect to the extent of impairment, errors in the measurement may arise from inconsistencies associated with interviewers, raters, and other unidentified sources. G theory incorporates these potential sources of error into a measurement model and estimates the components of variance associated with each source of variation in the  $20 \times 2 \times 2$  (S x I x R) design. Table 2 enumerates the sources of variation and presents the estimated variance components for the anxiety subscale. Three estimated variance components are large relative to other components. The first, for survivors, is analogous to true score variance in classical test theory and is expected to be large. The second, the survivor by interviewer interaction, represents one source of measurement error and is due to inconsistencies of the two interviewers in obtaining information for different survivors. The third is the residual term representing unidentified sources of measurement error. The estimated generalizability coefficient,  $\hat{\rho}^2$ , is 0.56 using two interviewers and two raters.

Table 2

GENERALIZABILITY OF MEASURES OF PSYCHIATRIC  
IMPAIRMENT OF DISASTER SURVIVORS

Source of Variation	Estimated Variance Component
Survivors (S)	1.84
Raters (R)	.21
Interviewers (I)	.49
SR	.27
SI	1.82
RI	.03
Residual (SRI,e)	1.58
Generalizability Coefficient ( $\hat{\rho}^2$ ) for one rater and one interviewer	0.33
Generalizability Coefficient ( $\hat{\rho}^2$ ) for two raters and two interviewers	0.56



## 2. THEORETICAL CONTRIBUTIONS

### 2.1. Estimated Variance Components: The Achilles Heel

Two major contributions of generalizability theory are its emphasis on the multiple sources of measurement error and its de-emphasis of the role played by summary reliability or generalizability coefficients. Estimated variance components are the basis for indexing the relative contribution of each source of error and the undependability of a measurement. Yet Cronbach et al. (1972) warned that the estimates of these variance components are unstable with usual sample sizes (cf. Lindquist, 1953; Smith, 1978; van der Kamp, 1976).

While we consider the problems associated with estimating variance components to be the Achilles heel of G theory, these problems afflict all sampling theories. One virtue of G theory is that it brings estimation problems to the fore and puts them up for examination. Estimation problems, then, are the Achilles heel of all theories that involve sampling.

With the importance and fallibility of estimated variance components so clearly recognized, we find it astonishing that so little attention has been given to this topic in the literature on G theory. Here we review the few studies that have been done and also point out research in the statistical literature which we hope will stimulate further work.

#### 2.1.1. Sampling variability of estimated variance components

Usually the estimate of a variance component ( $\hat{\sigma}^2$ ) is found using some linear combination of mean squares divided by a constant. The sampling variance of an estimated variance component ( $\hat{\sigma}^2$ ) is:

$$\begin{aligned}
 [11] \quad \text{VAR}(\hat{\sigma}^2) &= \frac{2}{c^2} \sum_q \text{VAR}(\text{MS}_q) \\
 &= \frac{2}{c^2} \sum_q \frac{\text{EMS}_q^2}{\text{df}_q},
 \end{aligned}$$

where  $c$  is the constant associated with the estimated variance component,  $\xi MS_q$  is the expected value of the mean square,  $MS_q$ , and  $df_q$  is the degrees of freedom associated with  $MS_q$ . (See Smith, 1978, for a lucid, terse development of this general formula.) For example, in Table 1, the variance component for persons is estimated by  $(MS_p - MS_{res})/n_i$ . With respect to [11],  $c$  refers to  $n_i$ ,  $MS_q$  refers to  $MS_p$ , and  $MS_{res}$ .

The problem of fallible estimates can be illustrated by expressing a mean square as a sum of population variances. In a two-facet, crossed ( $p \times i \times j$ ), random model design, the variance of the estimated variance component for persons--of the estimated universe score variance--is (Smith, 1978, Fig. 1):

$$[12] \quad \text{VAR}(\hat{\sigma}_p^2) = \frac{2}{(n_p-1)} \left( \sigma_p^2 + \frac{\sigma_{pj}^2}{n_j} + \frac{\sigma_{pi}^2}{n_i} + \frac{\sigma_{res}^2}{n_i n_j} \right)^2 + \frac{1}{(n_j-1)} \left( \frac{\sigma_{pj}^2}{n_i} + \frac{\sigma_{res}^2}{n_i n_j} \right)^2 \\ + \frac{1}{(n_i-1)} \left( \frac{\sigma_{pi}^2}{n_i} + \frac{\sigma_{res}^2}{n_i n_j} \right)^2 + \frac{1}{(n_i-1)(n_j-1)} \left( \frac{\sigma_{res}^2}{n_i n_j} \right)^2$$

With all of the components entering the variance of the estimated universe score variance,  $\hat{\sigma}_p^2$ , the fallibility of such an estimate is quite apparent (if  $n_i$  and  $n_j$  are modest). In contrast, the variance of the estimated residual variance ( $\hat{\sigma}_{res}^2$ ) has only one variance component.

$$[13] \quad \text{VAR}(\hat{\sigma}_{res}^2) = \frac{2}{(n_p-1)(n_i-1)(n_j-1)} \sigma_{res}^4.$$

In a crossed design, then, the number of components and hence the variance of the estimator increases from the highest order interaction component to the main effect components. Consequently, sample estimates of the universe-score variance--estimates of crucial importance to the dependability of a measurement--may reasonably be expected to be less stable than estimates of components of error variance.

The sampling variability of estimated variance components leads to a bandwidth-fidelity dilemma. Perhaps the greatest contribution of G theory is its applicability to complex, realistic, multifaceted

measurement designs. Hence, G theory provides great bandwidth in examining the dependability of behavioral measurements. However, for complex multifaceted measurements the variability of the estimated variance components is, in general, expected to increase. Hence, fidelity in estimation is lower for multifaceted universes than for single faceted universes. The bandwidth-fidelity dilemma is a dilemma associated with all applied statistics, not just with G theory.

### 2.1.2. Negative estimates of variance components

Negative estimates of variance components can arise because of sampling errors or because of model misspecification (Hill, 1970). With respect to sampling error, the one-way ANOVA illustrates how negative estimates can arise. The expected mean squares are:

$$[14] \quad \begin{aligned} \xi MS_W &= \sigma_1^2 \text{ and} \\ \xi MS_B &= \sigma_1^2 + n\sigma_2^2 = \sigma_{12}^2, \end{aligned}$$

where  $\xi MS_W$  is the expected value of the mean square within groups and  $\xi MS_B$  is the expected value of the mean square between groups. Estimation of the variance components is accomplished by equating the observed mean squares with their expected values and solving the linear equations (see Brennan, 1978a, for algorithms). If  $MS_W$  is larger than  $MS_B$ , the estimate of  $\sigma_2^2$  will be negative.

Realizing this problem in G theory, Cronbach et al. (1972, p. 57) suggested that "a plausible solution is to substitute zero for the negative estimate, and carry this zero forward as the estimate of the component when it enters any equation higher in the table of mean squares." [See Davis (1974, pp. 15-17) for a summary of approaches to treating negative estimates.] Notice that by setting negative estimates of variance components to zero, the researcher is stating that a reduced model provides an adequate representation of the data, thereby admitting that the original model was misspecified. Scheffe (1959), among others, has pointed out that while this is a reasonable solution to the problem, the sampling distribution of the (once negative) variance component as

well as those variance components whose calculation includes this component is more complicated and the modified estimates are biased. The problem of handling negative variance components is exacerbated in G studies with many crossed facets as [12] suggests.

A Bayesian approach to estimating variance components appears to be an attractive alternative to that of traditional sampling theory because it: (a) provides a solution to the problem of negative point and interval estimates of variance components; and (b) provides interval estimates of variance components interpretable with respect to the sample data, not to repeated sampling. While the Bayesian methods have rarely been applied to generalizability theory, the work of Box and Tiao (see also Davis, 1974; Fyans, 1977; Novick & Jackson, 1974; Novick, Jackson & Thayer, 1971) provides an important starting point.

Hill (1965, 1967, 1970; see also Novick et al., 1971) pointed out that sampling theory's usual unbiased estimate of  $\sigma_1^2$  ignores relevant information contained in  $\sigma_{12}^2 = \sigma_1^2 + n\sigma_2^2$ . For Bayesians, a negative estimate of the between persons component indicates that  $MS_W$  and  $MS_B$  are providing conflicting information. A Bayesian approach, then, incorporates information about  $\sigma_1^2$  that is constrained in both  $MS_W$  and  $MS_B$ . The approach also includes the constraint that  $MS_B \geq MS_W$  (Box & Tiao, 1973, see p. 254).

Fyans (1977) provided a general strategy for obtaining Bayesian estimates of the modes of the posterior distributions for variance components from crossed, partially nested, and completely nested designs. Following Box and Tiao's (1973) formulation, he assumed a locally uniform prior with  $p(\sigma) = \sigma^{-1}$  and constrained the variance components to be greater than or equal to zero. Under normality and independence, the joint mode ( $\hat{V}$ ) for the posterior distribution of any source of variation in a design is given by its sum-of-squares divided by the appropriate degrees of freedom plus two (Fyans, 1977). In a  $p \times i$  design, for example, the joint modes would be:

$$[15] \quad \begin{aligned} \hat{V}_p &= SS_p / (df_p + 2) , \\ \hat{V}_i &= SS_i / (df_i + 2) , \end{aligned}$$

and

$$\hat{V}_{pi} = SS_{pi} / (df_{pi} + 2) .$$

By equating the  $\hat{V}$ s--i.e., the adjusted mean squares--to their expectation, Bayesian modal estimates of the variance components ( $\hat{\sigma}^2$ ) can be obtained:

$$[16] \quad \hat{\sigma}_p^2 = (\hat{V}_p - \hat{V}_{pi}) / n_i$$

$$\hat{\sigma}_i^2 = (\hat{V}_i - \hat{V}_{pi}) / n_p$$

$$\hat{\sigma}_{pi}^2 = \hat{V}_{pi} .$$

Box and Tiao (1973) provided posterior distributions for variance components in crossed and completely nested designs. Fyans (1977) provided a posterior distribution for variance components for a partially nested design. With estimates in hand, a Bayesian generalizability coefficient can be obtained in the usual manner.

If the posterior modal values do not satisfy the constraint of non-negative estimated variance components, the Bayesian approach sets all joint modes equal to each other and uses a pooled estimate for the common value (Fyans, 1977, p. 151).

Finally, the interpretation of the Bayesian interval is more straightforward than is the interval obtained by sampling theory. In sampling theory, the probability statement refers to all possible confidence intervals rather than to a particular interval--the one in hand--that was constructed from sample data. In contrast, the probability statement associated with the Bayesian interval refers directly to likely values of the population variance component and not to replications of the design. In practical applications of any measurement theory, we make decisions on the basis of the sample data. Hence, the Bayesian interpretation corresponds to how variance components are used in practical applications of G theory.

Cronbach et al.'s (1972) evaluation of the potential contribution of Bayesian G theory has held up over time:

Bayesian statistical inference needs to be exploited systematically. It appears likely that developments now available in the statistical literature could, in some problems, profitably replace the methods of estimating variance components...[and universe scores]. Also, whereas we obtain all estimates from the G study, one could, by Bayesian methods, take into account the additional information offered by the D study to reach final conclusions about the generalizability of the D data [p. 336, *italics ours*].

Novick (1976) went even further than Cronbach et al., by pointing out that if one accepts de Finetti's exchangeability concept (see Section 2.2), "Generalizability Theory...is Bayesian in everything but a formal sense, though I do not think the authors see it that way" (p. 24). In our opinion, it is time that the formal sense in which G theory is Bayesian be systematically explored.

#### 2.1.3. Allocation of observations to reduce the sampling variability of estimated variance components

Woodward and Joe (1973) and Smith (1978) addressed the problem of how to allocate measurements (e.g.,  $n_i$  and  $n_j$ ) to maximize the generalizability coefficient (Woodward & Joe) or to produce the most stable estimates of variance components (Smith). They arrived at similar recommendations. In a  $p \times i \times j$  design, for example, as  $\hat{\sigma}_{res}^2$  increases relative to  $\hat{\sigma}_{pi}^2$  and  $\hat{\sigma}_{pj}^2$ , the optimal solution tends toward  $n_i = n_j$ . As  $\hat{\sigma}_{res}^2$  decreases relative to  $\hat{\sigma}_{pi}^2$  and  $\hat{\sigma}_{pj}^2$ , the optimal solution is to make  $n_i$  and  $n_j$  proportional to  $\hat{\sigma}_{pi}^2 / \hat{\sigma}_{pj}^2$ .

#### 2.1.4. Monte Carlo studies of variance components

Smith conducted Monte Carlo studies with: (a) three designs-- Design A =  $p \times i \times j$ , Design B =  $p \times (j:i)$ , and Design C =  $(j:p) \times i$ ; (b)  $n_p = 25, 50, 100$ ; (c)  $n_i$  and  $n_j = 2, 4, 8$ ; (d)  $\sigma_i^2 : \sigma_j^2 = 1:4, 4:1$ ; and (e)  $\sigma_{res}^2 = 20, 76$ . A random effects,  $p \times i \times j$  ANOVA model (assuming

additivity and independence) was used to generate normally distributed, integer scores. For each case, 300-500 replications were obtained.

Smith concluded that the estimated variance components from multifaceted generalizability studies contain sizable error. More specifically, he found that: (a) The sampling errors of variance components are much greater for multifaceted universes than for single faceted universes. (b) For  $\sigma_p^2$ , the sampling errors were large unless the total number of observations ( $n_{pij}$ ) was at least 800. (c) Stable estimates of  $\sigma_i^2$  and  $\sigma_j^2$  required at least eight levels of each facet. And (d) some nested designs produced more stable estimates than did crossed designs.

Simulations similar to those of Smith conducted by Calkins, Erlich, Marston, and Malitz (1978), Leone and Nelson (1966) and Smith (1980) have reached similar conclusions. Finally, Calkins et al. (1978) and Leone and Nelson (1966) also found many negative estimates of variance components, especially when the number of levels of a facet was small (e.g., five).

Recognizing these problems with estimated variance components, Smith (1978, 1980) proposed the use of several small G studies with many levels of one or a few facets instead of one large, crossed G study. One small G study might estimate  $\sigma_p^2$ ,  $\sigma_i^2$ , and  $\sigma_{pi}^2$ , while another might estimate  $\sigma_p^2$ ,  $\sigma_j^2$ , and  $\sigma_{pj}^2$  such that the total number of observations would be equal to that of the one large, crossed G study. We question, however, whether a universe of admissible observations represented in a series of G studies with more restricted universes is equivalent to the universe represented by one, larger universe. We also question whether the construction of a large G study from a series of smaller G studies would provide information appropriate for the decisionmaker's universe of generalization in a D study.

#### 2.1.5. Unbalanced Designs

An unbalanced design has unequal numbers of observations in its subclassifications. Two examples of unbalanced G study designs are (1) pupils nested in classes where class size is not constant, and (2) observers nested in occasions where unequal numbers of observers are present at each occasion.

In a comprehensive review, Searle (1971, p. 35) pointed out that the usual ANOVA approach to estimating variance components with balanced data--setting mean squares equal to their expectation--is not as straightforward when applied to unbalanced data. This section shows how the usual ANOVA approach to estimating variance components in balanced designs can be applied to unbalanced designs and points to an alternative approach using the computer (cf. Llabre, 1978, 1980; Brennan, Jarjoura & Deaton, 1980).

Several aspects of the ANOVA approach differ in unbalanced designs and are problematic. First, the sums of squares are not additive. The mean squares, therefore, may be unadjusted or adjusted for one or more effects. And the choice of adjustment is not always clear (see Searle, 1971). Second, solutions to the problem of non-additive sums of squares lead to biased estimation in mixed models. And, third, the simple rules for deriving expected values of mean squares (Cornfield & Tukey, 1956) do not apply to unbalanced designs. The coefficients in the expected mean square equations are algebraically and computationally complex.

Henderson (1953) proposed variations in the analysis of variance approach to deal with the first problem. The problem of biased estimation in mixed models is not a problem in G theory, since G theory is essentially a random effects theory. That is, G theory averages over fixed facets in a mixed model and so estimates only variances of random effects. Finally, computational complexity is reduced by using Rao's (1971, 1972) approach called minimum variance quadratic unbiased estimation (MIVQUE; see Llabre, 1978, 1980; Brennan, Jarjoura & Deaton, 1980). Incidentally, MIVQUE also avoids the problem of the order of the components.

MIVQUE is available in the VARCOMP procedure in the SAS (Statistical Analysis System, 1979) computer system to small designs. Brennan, Jarjoura and Deaton (1980) reviewed this and other computer programs for estimating variance components in unbalanced designs. They mentioned the limited storage capacities of many of the programs, which restrict their use to small designs. The major problem remaining in the estimating of variance components with unbalanced data, then, is to develop efficient computer programs that will estimate variance



componets in large generalizability designs without requiring prohibitively large storage capacities.

## 2.2. Fixed Facets

In G theory a fixed facet has a fixed set of conditions that appear in the G and D study. Although this definition parallels that given for a fixed factor in sampling theory, it also describes facets which are often considered in practice to be random. Loevinger (1965) goes so far as to argue that all facets must be considered fixed in any measurement theory. This issue is taken up below.

Statistically, G theory treats fixed facets by averaging (or summing) over the conditions of the fixed facet and examining the generalizability of these averages over the random facets (Cronbach et al., 1972, p. 60; see Erlich & Shavelson, 1976b, for a proof). While this treatment of fixed facets is justifiable on sampling theory grounds, it does not always lead to a sensible treatment of fixed facets in the measurement theory. This issue is also considered below.

### 2.2.1. Fixed versus random facets

Often a test developer has a fixed set of items which he considers to be a random sample of items from some more or less well defined universe. Can this set of items legitimately be considered a random sample or, as Loevinger (1965) insists, should it be considered a fixed facet?

The objection to all psychometric developments that assume random sampling of items or tests is in the first instance that they grossly misrepresent the actual case, which is almost invariably expert selection rather than random sampling. But there is a subtler and deeper point. The term population implies that, in principle one can catalog, or display, or index all possible members, even though the population and the catalogue [sic] cannot be completed. Statistical sampling must be tied to such a display and indexing system, else it cannot be random [Loevinger, 1965, p. 147].

One possible way to resolve this problem is suggested by de Finetti's (1964) concept of exchangeability (cf. Kingman, 1978;

Lindley & Novick, 1979; Novick, 1976; see also Davis, 1974; Fyans, 1977). Put (perhaps, too) simply, this concept states that even though conditions of a facet have not been sampled randomly, the facet may be considered to be random if conditions not observed in the G study can be exchanged with the observed conditions. Formally,

The random variables  $X_1, \dots, X_n$  are exchangeable if the  $n!$  permutations  $(X_{k1}, \dots, X_{kn})$  have the same  $n$ -dimensional probability distribution. The variables of an infinite sequence  $X_n$  are exchangeable if  $X_1, \dots, X_n$  are exchangeable for each  $n$  [Feller, 1966, p. 225].

Although not explicitly stated, the concept of exchangeability is evident in discussions of G theory applications. For example, Elffers and Tavecchio (1979, p. 5) argued that as long as the conditions of the facets in the G study are not very different from the larger set, the facet can be considered random.

Viewed from the exchangeability perspective, the issue of fixed or random facets is not whether one can catalog (etc.) all possible members of a population, but whether the members in hand are exchangeable with other potential members. In terms of item sampling, if one set of persons and items to which  $\rho^2$  is generalizable is a set of such persons and items jointly exchangeable with the present sample, it is reasonable to consider the item facet to be random.

The concept of exchangeability, at a minimum, provides reasonable grounds for considering whether a facet is random or fixed. At best, it suggests that random facets abound.

### 2.2.2. Statistical treatment of fixed facets

G theory treats a fixed facet in one of several different ways. Perhaps the most frequent approach is to average scores over the conditions of the fixed facet and examine the generalizability of this average over the random facets. For example, general aptitude batteries like the ACT Assessment are designed to predict future academic achievement. Clearly, the ACT subtests are fixed, so scores would be averaged

over subtests and the generalizability of this average examined over the random facets of the design. In this case, averaging scores over subject matters makes good sense from the standpoint of prediction of academic success. Notice that by averaging over the conditions of a fixed test, G theory is essentially a random model theory.

A second approach is to examine the generalizability of scores at each condition of the fixed facet. For example, the generalizability of scores on the ACT Assessment would be examined separately for each subtest. Often, in treating each condition of a fixed facet separately the decisionmaker is willing to consider the conditions of the facet as a profile of scores. Hence, the analysis focuses on estimating each subject's universe score on each condition of the fixed facet (Cronbach et al., 1972). For example, the ACT subtests might be considered a profile and estimates of students' universe scores would be obtained for each subtest. (For a discussion of this approach, see Section 2.5.)

The third approach is to choose one of the first two approaches on the basis of the estimated variance of the conditions of a fixed facet. In other words, in the absence of strong a priori reasons for averaging over the conditions of a fixed facet or treating the conditions as a profile, the decisionmaker examines the variability of the conditions of a fixed facet. If the variability is minimal, the scores may be averaged over the conditions of the fixed facet. If the variability is substantial, the decisionmaker may choose to treat each condition separately or consider the conditions as a profile.

The decision about how to treat a fixed facet in a G study is not necessarily straightforward, as illustrated by a study of measures of teacher behavior (Erlich & Shavelson, 1978). Teachers were observed on three occasions by two raters while teaching reading and math. Subject matter was treated as fixed, and teachers' scores were averaged over reading and math lessons. Teaching behavior, however, is quite different during reading and math. Averaging over those two subject matters may have distorted the phenomena being observed as well as the estimated universe score variance. A preferable strategy in this study might have been to examine the generalizability of the reading and math data separately or as a profile. However, an elementary school principal

might be interested in a teacher's behavior in general and so be quite willing to use an average over reading, math, and other subjects. For the principal's purpose, Erlich and Shavelson's (1978) treatment of the subject matter facet may have been appropriate.

Since the decisionmaker determines whether a facet is fixed or random, the most reasonable solution may be to report, in a G study, the variance components and generalizability coefficients for each condition of a fixed facet separately (usually there are only a few conditions of a fixed facet) and in combination (averaged over the conditions). By doing so, the decisionmaker can choose which data are most pertinent to a proposed D study.

### 2.3. Criterion-referenced Measurement

The term criterion-referenced measurement (CRM) has been defined in a variety of ways.<sup>2</sup> Most of these definitions include a well-specified content domain (Hambleton & Novick, 1973; Popham, 1975). Following Cronbach et al. (1972) and Brennan (1980), we speak of interpretations of test scores as being criterion, or content, or domain referenced. The observed score is interpreted as being representative of the universe of content from which it is sampled.

Since criterion-referenced interpretations consider an individual's status independent of other persons (cf. absolute decisions in Cronbach et al., 1972, p. 14), "The primary question is: how large is the error arising from incomplete observation?" (Cronbach et al., 1972, p. 23). The errors,  $\Delta$  and  $\epsilon$ , and the G coefficient for absolute decisions speak directly to criterion-referenced interpretations (cf. Brennan, 1980; Brennan & Kane, 1977a,b; Hambleton, Swaminathan, Algina & Courson, 1978; Kane & Brennan, 1980; Linn, 1979; Shavelson, 1979).

Cronbach et al. (1972) discussed three approaches to estimating universe scores: regression, interval, and Bayesian estimates. Generalizability theory greatly complicates regression estimates "when it recognizes that conditions may not be equivalent and considers any set of conditions to be a sample from a universe" (Cronbach et al., 1972, p. 140). Moreover, with interval estimates, the probability statements about nonrandomly selected individuals are not justifiable. Finally, Cronbach et al. (1972)

suggested that estimation problems might be solved with a Bayesian approach (see Section 2.1.2) if this approach could be extended to complex designs.

In reviewing the literature, we were surprised to find that the call for systematic work on Bayesian estimates of universe scores in G theory has not been answered. Fyans (1977) provided a good summary of methods proposed by Novick (1969; Novick & Hall, 1965), Box and Tiao (1968), and Lindley (1971). The work of Novick (Novick, Jackson, Thayer & Cole, 1972; Novick et al., 1971; Novick, Lewis & Jackson, 1973; Novick, Jackson & Thayer, 1971; Lewis, Wang & Novick, 1975; see Molenaar & Lewis, 1979, for a computer program) and others on m-group regression and the work of Wilcox (1978) with empirical Bayes estimation procedures for true scores in the compound binomial error model seem to be a logical starting place for Bayesian estimation of universe scores.

While G theory focuses on estimation, most of the CRM literature has focused on generalizability coefficients (cf. Brennan, 1980; Brennan & Kane, 1977a,b; Kane & Brennan, 1980). Brennan and Kane (1977a) proposed a coefficient which paralleled Livingston's (1972a) coefficient developed within classical test theory. The reasoning went as follows. In CRM, we are interested in the difference between a person's universe score in a well-defined behavioral domain and some criterion in that domain. In estimating a person's universe score from his observed score, the error is

$$\Delta_{pI} = (X_{pI} - \lambda) - (\mu_p - \lambda) = X_{pI} - \mu_p,$$

where  $\lambda$  is the criterion. From this formulation, their index of dependability for a domain-referenced test-- $\phi(\lambda)$ --is:

$$\begin{aligned} [17] \quad \phi(\lambda) &= \frac{\xi_p (\mu_p - \lambda)^2}{\xi_I \xi_p (X_{pI} - \lambda)^2} \\ &= \frac{\sigma_p^2 + (\mu - \lambda)^2}{\sigma_p^2 + (\mu - \lambda)^2 + \sigma_i^2/n_i + \sigma_{pi}^2/n_i} \end{aligned}$$

$$[18] \quad = \frac{\sigma_p^2 + (\mu - \lambda)^2}{\sigma_p^2 + (\mu - \lambda)^2 + \sigma_\Delta^2} \cdot$$

In the special case where  $\lambda = \mu$ , the index of dependability is equal to the G coefficient for absolute decisions:

$$[18a] \quad \rho^2 = \phi = \frac{\sigma_p^2}{\sigma_p^2 + \sigma_\Delta^2} \cdot$$

Brennan (1980) characterized  $\phi(\lambda)$  as follows: (a) It uses the error associated with absolute decisions ( $\sigma_\Delta^2$ ) rather than the error for relative decisions ( $\sigma_\delta^2$ ). (b) It varies from 0 to 1. (c) It varies with different values of  $\lambda$ , i.e., the numerator depends on the universe score variance and the squared distance of the population mean from the criterion. (For a critique of this characteristic, see Harris, 1972; Linn, 1979; Shavelson, Block & Ravitch, 1972; but see also Kane & Brennan, 1980; Livingston, 1972b.) And (d)  $\phi(\lambda)$  can be positive with zero universe-score variance. Brennan (1979b, pp. 27-28; see also Brennan, 1980; Kane & Brennan, 1980) distinguished the interpretation of  $\phi(\lambda)$  from that of  $\phi$  as follows: " $\phi(\lambda)$  provides an estimate of the dependability of the decisions based on the testing procedure [including chance agreement in scores];  $\phi$  provides an estimate of the contribution of the testing procedure to the dependability of such decisions."

In generalizability theory, we are primarily interested in variance components rather than generalizability coefficients. And estimates of variance components are not changed by introducing a criterion for the purpose of estimating an index of dependability (Linn, 1979).

If, in CRM, interest attaches to a mastery-nonmastery decision, as implied by  $\phi(\lambda)$ , a coefficient of generalizability seems less important than an estimate of the probabilities of false positive ( $\alpha$ ) and false negative ( $\beta$ ) decisions. The distinction between  $\alpha$  and  $\beta$  is important because the seriousness of each may not be the same, which, in turn, may affect the decision rule used to determine whether  $\mu_p$  is above or below  $\mu_o$ , the criterion score.

Wilcox's (1977, 1979) work provides a starting point for dealing with this situation. Assume that  $n_1$  items are randomly sampled from a skill domain. They are administered to an examinee in order to determine whether his true score,  $\mu_p$ , is above or below the known criterion score  $\lambda$ . If  $\mu_p \geq \lambda$ , the examinee is a master. However, the decision about  $\mu_p \geq \lambda$  is made if and only if the examinee's observed score,  $X$ , is greater than or equal to  $X_o$ , the "operational criterion" for deciding mastery. Note that the choice of  $X_o$  may incorporate the decisionmaker's notion of the losses associated with  $\alpha$  and  $\beta$  (cf. Wilcox, 1979, p. 60) and so  $X_o$  may not equal  $\lambda$ . In order to estimate  $\alpha$  and  $\beta$ , Wilcox (1977) used an empirical Bayes approach assuming either a beta distribution or a normal distribution (on transformed  $X$ 's) for true scores and a binomial probability function of observed scores given true scores. Noting problems with beta and normal priors, Wilcox (1979) worked out upper and lower bounds for  $\alpha$  and  $\beta$  which make no assumptions about the form of the true score distribution.

#### 2.4. Symmetry

The purpose of psychological measurement has typically been to differentiate individuals. The focus on individuals is reflected in Cronbach et al.'s mentioning only the case of measuring attributes of schools or classrooms (teachers) as an alternative to measuring attributes of individuals within them. Unlike the traditional concentration on individuals, however, Cardinet, Tourneur and Allal (1976a,b, in press; Cardinet & Tourneur, 1974, 1977; Tourneur, 1978; Tourneur & Cardinet, 1979) recognized that the focus of measurement may change depending on a particular decisionmaker's purpose. "One can easily cite cases, particularly in educational research, where the purpose of measurement is to compare the rates of success for different test items, or for different instructional treatments" (Cardinet et al., in press, p. 6; cf. Wood, 1976a). Individual differences, then, may represent a source of error--rather than universe score--variation in the measurement.

Cardinet and his colleagues speak of a principle of symmetry: "The principle of symmetry of the data is simply an affirmation that each of

the facets of a factorial design can be selected as an object of study, and that the operations defined for one facet can be transposed in the study of another facet" (Cardinet et al., in press, p. 7).

The principle of symmetry led Cardinet et al. (in press) to distinguish between four stages of a measurement study: (a) the observation design, (b) the estimation design, (c) the measurement design, and (d) the optimization design. By distinguishing these four stages, Cardinet et al. (in press) were able to disentangle measurement considerations (e.g., specification of the object of measurement) from the computations that yield estimates of variance components. "Until the two kinds of problems (ANOVA estimates, measurement) were clearly disentangled, no multi-purpose measurement was possible" (Cardinet, personal communication, May 9, 1980).

The first stage--observation--includes the choice of facets and conditions and computation of mean squares. The second stage--estimation--involves the decision about whether the facets are finite or infinite and random or fixed (cf. Wood, 1976a, for an application of finite, random facets), and the estimation of variance components. The third stage--measurement--specifies which facet (or combination of facets, see below) is the focus of measurement and which facets may limit the generalization of the measurement (i.e., sources of error). Estimates of error ( $\hat{\sigma}_\delta^2, \hat{\sigma}_\Delta^2$ ) and generalizability ( $\xi_{p^2}$ ) are obtained in this stage. The fourth stage provides information relevant to alternative D-study designs.

As an example of the four stages in a study, consider a G study of student evaluations of teaching. An observation design might have classrooms (c), students nested within classrooms (s:c), and items (i) crossed with both classrooms and students (see Smith, 1979, for a discussion of designs for student ratings). For simplicity, assume that the same number of students is observed in each classroom (but see Section 2.1.5).

In the estimation design, a decision is made about whether the model is random or fixed (see Section 2.2 for a discussion of fixed facets). The variance components, then, are estimated according to the appropriate model.



In the measurement design, the focus of measurement--classrooms-- is identified along with sources of error variation (students and items) and  $\sigma_{\delta}^2$  or  $\sigma_{\Delta}^2$  and  $\xi\rho^2$  are estimated (see Smith, 1979). Kane (& Brennan, 1977; Kane et al., 1976; see also Gillmore et al., 1978; Smith, 1979) pointed out that the magnitude of the generalizability coefficient depends upon whether items is considered a fixed or random facet. While the expected observed score variance remains the same--  $\sigma_c^2 + \sigma_{ci}^2/n_i + \sigma_{(s,s:c)}^2/n_s + \sigma_r^2/n_i n_s$ --the universe score variance is  $\sigma_c^2 + \sigma_{ci}^2/n_i$ . Kane and Brennan (1977; Kane et al., 1976) also discussed the case where the student facet is fixed and the item facet is random--i.e., the case of a nested universe. Finally, Kane (et al., 1976; see also Kane & Brennan, 1977) pointed out that the instructor effect is confounded with course content, differential selection of students into classes, observation occasion, and so on. This confounding will inflate the variance attributed to instructor to an unknown extent. Gillmore (1980; et al., 1978) and Smith (1979) presented designs (and data) that reduce this confounding.

Incidentally, Kane and Brennan (1977) related generalizability theory's approach to estimating the reliability of classroom means to approaches proposed in classical theory. They showed that classical theory approaches treated one facet as random while the other facet was implicitly treated as fixed.

Finally, the fourth stage of the measurement study would consider alternative sample sizes for students, items or both (depending on whether the model is random or mixed). It would also take into account the possibility of nesting items within students (see Cronbach et al., 1972, on matrix sampling studies, p. 214ff).

The principle of symmetry leads to the possibility of multifaceted populations. Cardinet & Tourneur (1977; with Allal, 1976 (a,b), in press) noted that, in surveys of educational achievement, the focus of measurement is on activities (objectives), years or schools, and not on students. In their example, the survey might have focused on the attainment of educational objectives (o) nested in content units (o:c) and crossed with students (s). The universe score of interest, then, is that of (say) objectives nested within content units (o:c) while

the sampling of students gives rise to errors of measurement. [A variety of such designs is given in Cardinet, Tourneur & Allal (1976; in press.)]

The notions of symmetry and multifaceted populations lead to several important consequences. In the (o:c) x s design described above, for example, an increase in the generalizability coefficient would be obtained by increasing sample size (e.g., number of students). This is just what would happen in sampling theory by specifying power,  $\alpha$ , a difference in means to be detected, and then calculating n. In short, both approaches lead to a decrease in the standard error of the mean. This is one point where the specialized area of measurement theory meets the more general area of sampling theory's estimation of differences between means.

A second consequence of symmetry and multifaceted populations is that the decisionmaker is able to systematically examine the assumption that measures are taken on a sample of persons from a homogeneous population. For example, if a population consists of subjects nested within sex and socioeconomic status (SES), variance components can be estimated for each facet. If the variance components for sex and SES are negligible for the particular attribute being measured, the decisionmaker can assume a homogeneous population and so reduce the design of the D study. If the components are sizeable, the decisionmaker may calculate separate estimates for each subgroup.

One final contribution of the principle of symmetry to be mentioned here is that it has led Cardinet and his colleagues to consider that case of a fixed facet comprising the focus of measurement. For example, evaluation of teachers in a school system might involve observers periodically observing the teachers. In this case, estimates of differences between a fixed set of teachers or, more appropriately, estimates of their universe scores might be the focus of measurement. Likewise, in industrial settings where supervisors' ratings of employees are gathered, one might consider employees as a fixed facet within some period of time.

### 2.5. Multivariate Generalizability

Educational and psychological measurements often involve multiple scores describing individuals' aptitudes or skills. For example, composites of CTBS subtest scores are used in classifying children for educational programs; and university instructors in science laboratories look for proficiency in students' manipulative, observational, interpretative, and planning skills (cf. Wood, 1976). Although multiple scores may be conceived as vectors, and thus should be treated simultaneously, the majority of generalizability and decision studies have not done so. Rather, each variable has been treated separately. One reason is the paucity of theory and procedures for examining multiple outcomes. Another reason is that the multivariate literature is not always easily comprehended. In an attempt to remedy this situation, we outline Cronbach et al.'s (1972) contributions more extensively here than in other parts of the review. We then describe further developments in multivariate generalizability and point to problems still in need of attention.

#### 2.5.1. Background

In extending the notion of multifaceted error variance to multivariate designs, Cronbach et al. (1972) stressed the separate treatment of the scores rather than the use of a composite of them. This permits the decisionmaker to examine variances of and covariances among the variables, and to formulate an optimal D-study design. As was the case in the development of univariate G theory, Cronbach et al. focused on methods of obtaining and interpreting variance components. Multivariate G theory decomposes both variances and covariances into components, whereas univariate G theory examines only components of variance. The expected mean product equations are solved in analagous fashion to their univariate counterparts (for an elementary exposition, see Travers, 1969). For example, the decomposition of the variance-covariance matrix in a one-facet, crossed design with two dependent variables is:

$$\begin{aligned}
 [19] \quad & \begin{bmatrix} \sigma^2(1^X_{pi}) & \sigma(1^X_{pi}, 2^X_{pg}) \\ \sigma(1^X_{pi}, 2^X_{pg}) & \sigma^2(2^X_{pg}) \end{bmatrix} = \begin{bmatrix} \sigma^2(1^p) & \sigma(1^p, 2^p) \\ \sigma(1^p, 2^p) & \sigma^2(2^p) \end{bmatrix} \\
 & \quad \text{(observed scores)} \qquad \qquad \text{(persons)} \\
 & \qquad \qquad \qquad + \begin{bmatrix} \sigma^2(1^i) & \sigma(1^i, 2^g) \\ \sigma(1^i, 2^g) & \sigma^2(2^g) \end{bmatrix} \\
 & \qquad \qquad \qquad \qquad \qquad \text{(conditions)} \\
 & \qquad \qquad \qquad + \begin{bmatrix} \sigma^2(1^{pi,e}) & \sigma(1^{pi,e}, 2^{pg,e}) \\ \sigma(1^{pi,e}, 2^{pg,e}) & \sigma^2(2^{pg,e}) \end{bmatrix} \\
 & \qquad \qquad \qquad \qquad \qquad \text{(residual)}
 \end{aligned}$$

where  $1^X_{pi}$  = score of variable 1 for person  $p$  observed under condition  $i$ ,  
 $2^X_{pg}$  = score on variable 2 for person  $p$  observed under condition  $g$ , and  
 $1^p$  = abbreviated for  $1^\mu_p$ : the universe score on variable 1 for person  $p$ .

In [19], the term  $\sigma(1^p, 2^p)$  is the covariance between universe scores on variables 1 and 2, say, ratings on two aspects of writing: organization and coherence. The term  $\sigma(1^i, 2^g)$  is the covariance between scores on the two variables due to the conditions of observation. Facet  $i$  may be the same as facet  $g$ , for example, when the same essay is used to obtain ratings of organization and coherence.

An important aspect of the development is the distinction between linked and unlinked conditions. When conditions for observing different variables are selected independently, the expected values of all

error components of covariance are zero. With independent sampling, then, the expected values of the off-diagonal elements in the last two matrices of [19] are zero. For example, in the illustration above, if the essays used to obtain ratings of organization are selected independently of the essays used to obtain ratings of coherence, then the expected value of the component of covariance  $\sigma_{(1,2)g}$  is zero. When conditions for observing multiple outcomes are not selected independently, but are jointly sampled, the expected values of all components of covariance are non-zero. Cronbach et al. (1972) presented the following example of a non-zero component of covariance for conditions.

Suppose that the design calls for teacher  $i$  to rate pupils  $p$  on both ability  $v_1$  and motivation  $v_2$ . Some teachers give higher ratings on the average than other teachers do; the  $i$  component represents this bias. The constant errors in  $v_1$  ratings are likely to covary (over teachers) with the constant errors in  $v_2$  ratings. The covariance  $\bullet\sigma_{(1,2)i}$  then would be positive. [Page 277; following Cronbach et al., the bullet symbol ( $\bullet$ ) indicates linkage.]

The literature reviewed below, while acknowledging the possibility of linked conditions, addresses only the unlinked case.

In their discussion and illustrations of multivariate generalizability analysis, Cronbach et al. (1972) did not develop a multivariate generalizability coefficient, but focused almost entirely on the interpretation of components of variance and covariance. They examined components to rule out "distressing" counterhypotheses. In an analysis of verbal and performance scores from the WISC and WAIS, for example, test forms were linked because verbal and performance scores were observed on the same form (WISC or WAIS) on the same day. However, the small variance and covariance components reported for forms indicates that linkage is not problematic here.

Travers (1969) developed a correction for attenuation analogous to Spearman's classical formula,  $r_{T_X T_Y} = r_{XY} / \sqrt{r_{XX'} r_{YY'}}$ . Travers

(1969, p. 344ff. and Cronbach et al., 1972, p. 287) showed that Spearman's formula can be restated as

$$[20] \quad \hat{\sigma}(\mu_{1p}, \mu_{2p}) = \frac{\hat{\sigma}(\mu_{1p}, \mu_{2p})}{\sqrt{\hat{\sigma}^2(\mu_{1p}) \cdot \hat{\sigma}^2(\mu_{2p})}} .$$

Whether the expected covariance between observed scores equals the covariance between universe scores depends upon linkages among conditions of facets in the design. In a two-facet design, for example, when  $i$  and  $g$  are independently sampled, the expected covariance of  ${}_1X_{pi}$  with  ${}_2X_{pg}$  is  $\sigma(\mu_{1p}, \mu_{2p})$ . When  $i$  and  $g$  are linked, however, the expected covariance is  $\sigma(\mu_{1p}, \mu_{2p}) + \sigma(\mu_{1pi}, \mu_{2pg})$ . When the expected covariance between observed scores is used as an estimate of the covariance between universe scores, then, the corrected correlation obtained with joint sampling will tend to be higher than that obtained with independent sampling, although the effect decreases as the number of levels of the  $i$ -facet increases.

Cronbach et al. (1972) also developed a multivariate predictor of the universe score. In the univariate case, the universe score is estimated from the regression of the universe score on observed scores (p. 103):

$$[21] \quad \hat{\mu}_p = (\hat{\xi}\rho^2)X_{pI} + (1 - \hat{\xi}\rho^2)X_{pI}$$

In the multivariate case, the regression equation for a particular dependent variable includes not only the observed scores on that variable, but also observed scores for all other variables in the set. The multiple regression coefficients are estimated using linked or unlinked covariances, depending upon the anticipated design of the study.

The set of multiple regression equations produces a profile of estimated universe scores for each person. This profile is more reliable (and usually flatter) than that based on univariate regression

equations. In an example using data from the Differential Aptitude Tests (DAT), Cronbach et al. (1972) reported reductions in error variance as large as 42% when all subtests were used as predictors compared to error variances from single predictors! (They added, however, that the number of predictors used in each equation should be guided by the sample size,  $n_p$ . See Darlington, 1978, and Fyans, 1978, for regression procedures that yield reduced sampling errors of regression estimates when the number of predictors is large relative to the sample size.) The important finding for counseling and research is that observed profiles and those estimated from univariate regressions may be much farther from the true profiles than multivariate estimates.

Surprisingly little has been published on multivariate generalizability theory in recent years. A multivariate generalizability coefficient has been developed for the limited case where the decision-maker simply wants to maximize the generalizability of a composite. The following sections discuss this multivariate G coefficient, the interpretation of canonical variates in multivariate analyses, and the choice between univariate and multivariate analyses.

#### 2.5.2. Multivariate generalizability coefficient

Bock (1963, 1966; see also Haggard, 1958) and Conger and Lipshitz (1973; Conger, 1974) developed multivariate analogues of test reliability for one-facet designs. From a random-effects, multivariate analysis of variance of standardized scores, the multiple discriminant functions are determined so as to maximize the ratio of between-person variation to within-person variation. Since the Bock and Conger and Lipshitz coefficients do not differentiate between different sources of error variance, they have limited utility for the design of decision studies.

The only multivariate reliability coefficient anchored in generalizability theory was developed by Joe and Woodward (1976). Their approach distinguished between G and D studies and generalized the work of Bock and Conger and Lipshitz to a variety of multifaceted designs with crossed and nested facets.

For the two-facet, fully crossed design, the multivariate coefficient is

$$[22] \quad \rho^2 = \frac{\underline{a}' \underline{V}_p \underline{a}}{\frac{\underline{a}' \underline{V}_p \underline{a}}{n_i'} + \frac{\underline{a}' \underline{V}_{pi} \underline{a}}{n_j'} + \frac{\underline{a}' \underline{V}_{pj} \underline{a}}{n_i' n_j'} + \frac{\underline{a}' \underline{V}_e \underline{a}}{n_i' n_j'}}$$

where  $\underline{V}$  = a matrix of variance and covariance components estimated from mean square matrices,  
 $n_i'$  and  $n_j'$  = the number of conditions of facets  $i$  and  $j$  in a D study, and  
 $\underline{a}$  = the vector of canonical coefficients that maximizes the ratio of between-person to between-person plus within-person variance component matrices.

For one-facet designs with large samples and one condition of the facet, Joe and Woodward's coefficient is equivalent to the coefficients developed by Bock (1966) and Conger and Lipshitz (1973; see also Conger, 1974). The value of Joe and Woodward's approach is that it allows us to maximize a generalizability coefficient by assessing the magnitude of different sources of error and so design D studies that reduce the sources of large error variation.

One limitation of all of the above approaches arises when variance component matrices are not positive definite or positive semidefinite. Joe and Woodward (1976) recommended using "extreme caution" and suggested that negative definite matrices should not be used in the estimation of variance component matrices and generalizability coefficients. As expected, the problem with negative estimates of variance components in univariate generalizability extends to the multivariate case; solutions need to be worked out in both arenas (see Section 2.1).

### 2.5.3. Interpretation of canonical variates

There is a set of canonical coefficients ( $\underline{a}_s$ ) for each characteristic root ( $\lambda_s$ ) in [22]. Each set of canonical coefficients defines a



composite of the scores. By definition, the first composite is the most reliable while the last composite is the least reliable.

Attention should be paid to the interpretation of these composites (see, for example, Fyans, Salili, Maehr, & Desai, 1980; Harnqvist, 1973; Peng & Farr, 1976). Conger and Lipshitz (1973), for example, examined the canonical coefficients in an illustrative analysis of data from the WISC to provide further information about common diagnostic interpretations of differences among subscales. Since all subtests had positive weights on the first, most reliable composite, they interpreted this finding as providing support for the use of the total IQ score. The second most reliable composite was not the expected contrast between verbal and performance IQ. Rather, this composite was a contrast between the verbal subtests and the subtests of Block Design, Object Assembly, and Mazes. This contrast was more reliable than the verbal-performance contrast. The remaining canonical variates also provided unexpected contrasts among subtests.

In using the multivariate G coefficient, the data, not the investigators, define the composites of maximum generalizability. This empirically-derived coefficient may not correspond to the way composites are defined by theory (e.g., theory of human abilities) or practice (interpretation of subtests for classification of applicants). Rather, we would prefer to estimate the generalizability of a composite given a set of constraints. Two issues are involved: determining the weights and estimating the generalizability of the composite.

The weights can be determined using psychological theory or practical application. The weights might be a set of orthogonal coefficients. For example, Harnqvist (1973) examined canonical coefficients in an analysis of data from the Primary Mental Abilities (PMA) battery. He could have used orthogonal coefficients to weight the four ability subtests in his analysis to form a verbal-numerical contrast, hypothesized to be important in factor theories of intelligence.

A second method of obtaining weights conforming to theory or practice is to establish the weights using confirmatory maximum-likelihood factor analysis (Jöreskog, 1969). In an illustration of this method, Jöreskog analyzed scores of nine mental ability tests. In

one example solution, tests were hypothesized to measure three factors --visualization, verbal intelligence, and speed. The first model specified three factors with loadings reflecting the hypothesized structure. The resulting  $\chi^2$  statistic indicated a poor fit to the data. Jöreskog suggested several ways to determine the cause of the poor fit; all of them involve relaxing one or more restrictions in the model and re-evaluating the fit.

The problem with repeated testing of the model is exactly the same problem that led us to seek alternatives to maximizing the ratio of between-person to between-person plus within-person variation. Namely, the changes in the model are determined by the data, not by theory or practice. The investigator should stop modifying the coefficients before the resulting composites become uninterpretable, even if the fit to the data is poor.

These approaches to determining the weights of the variables in composites may involve a tradeoff between interpretability and generalizability. We emphasize interpretability; a composite that is generalizable but not interpretable will not be of much use.

With respect to estimating the generalizability of the composite formed either a priori or by an empirical test of goodness of fit, the most straightforward approach is a univariate rather than a multivariate analysis. The results of a univariate generalizability analysis would be identical to those of a multivariate generalizability analysis in which the weights of the composite define the a vector in [22].

#### 2.5.4. Relation between multivariate and univariate G theory

When multiple scores are conceived of as composites, a multivariate generalizability analysis is appropriate because it explicitly takes into account the covariation among the scores. By partitioning the variance-covariance matrix for observed scores into matrices of components of variance and covariance for universe scores and error, the investigator can identify major sources of error variation and covariation, essential information for designing an optimal D study. In terms of a generalizability coefficient, if the decisionmaker's interest is in obtaining a composite with maximum generalizability,

Joe and Woodward's generalizability coefficient is appropriate. If the decisionmaker wants to assess the generalizability of a composite defined by theory or practice, a univariate generalizability study of the composite will produce the same results as the multivariate analysis substituting the a priori weights into Joe and Woodward's multivariate formula.

The above discussion does not address the generalizability of a profile of scores. In a profile, interest lies in the pattern of scores, not in a composite of them. Cronbach et al. (1972) described in detail the estimation of universe scores in a profile. In their formulation, univariate generalizability coefficients of the scores in the profile serve as one basis for judging the generalizability of individual scores in the profile. They further suggested that the multivariate approach of Bock (1966)--and so by implication that of Joe and Woodward--can be used to reduce and reorganize the profile. The multivariate generalizability coefficients may show that some combinations of scores are measured more reliably than is needed, while others are not measured with sufficient precision for the decisionmaker's needs. The decisionmaker can use this information to eliminate scores in the profile or to design ways of obtaining more generalizable measures of them.

## 2.6. Sampling in Observational Measurement

The ability to estimate the contribution of multiple sources of error affecting observational measurements is one of the major contributions of G theory. However, a problem still to be resolved is how to allocate observations taking into account the linkage arising from adjacent observations.

The amount of observation time can vary on two dimensions: (a) facets that affect the number of observations, and (b) facets that affect the length of observation periods. The problem with developing procedures for estimating reliability for different numbers and lengths of observations is that observations may be correlated (called linked in the previous section; see also Cronbach & Furby, 1970, p. 69; Cronbach et al., 1972, p. 268ff). That measures obtained on the same

day may not be independent is intuitively reasonable. Not as obvious, perhaps, is the linkage among measures obtained on different occasions. The linkage, as in any time series problem, is a matter of degree. Measures of teacher behavior obtained at different times during a class period, for example, may agree more than measures obtained at different times during a school day, which are, in turn, likely to agree more than measures obtained on different days.

Different degrees of linkage can occur even within a short span of time, as is illustrated in Table 3. Webb (1980) observed the proportion of each minute that a student worked alone without communicating with other students or with the teacher--a variable commonly observed. Behavior of junior high school students was recorded during five-minute segments. To determine whether behavior during adjacent one-minute intervals within a segment was more similar than behavior from intervals that were further separated, the matrix of intercorrelations among the scores for one-minute intervals was calculated. The matrix exhibits a simplex pattern. Thus, different degrees of linkage are clearly evident, especially within a time period as short as five minutes.

Table 3  
CORRELATIONS OF STUDENT BEHAVIOR ("WORKS ALONE")  
CALCULATED FOR CONSECUTIVE ONE-MINUTE INTERVALS  
(N = 50)

Minute	2	3	4	5
1	.58	.43	.34	-.12
2		.56	.32	.10
3			.52	.06
4				.23

The simplex pattern just described extended to observations made at different times on one day and to observations made on different days. As expected, the correlations between observations drawn from

different five-minute segments of the same class hour were lower than for adjacent one-minute intervals. Correlations between one-minute observations on different days were lower still.

The only attempt thus far to estimate the effect on reliability of varying the length and number of observation periods has been made by Rowley (1976, 1978). Rowley (1978) described the effects of varying number and length of observations separately and simultaneously. Unfortunately, Rowley treated observation periods as if they were unlinked, and so applied the Spearman-Brown formula inappropriately.

The more general questions that need to be addressed in further research are (1) how can the correlation between observation periods occurring closer or further in time be taken into account, and (2) does the recognition of linked conditions of a facet make a difference in the theory or in the analysis?

A potential solution is the following. First, consider length as a vector of scores corresponding to different durations of observation time or consecutive observation intervals. This is tantamount to defining the universe of generalization to include multiple scores as well as multiple sources of measurement error. Thus, in Webb's (1980) observational study described above, the proportion of each minute spent working alone in the five-minute observation period may be entered as a vector of five scores. Next, examine the generalizability of observational measurements with a one-facet (number of observations) multivariate analysis of variance (see Section 2.5). This analysis would enable the decisionmaker to estimate the number of observations needed in a D study and the optimal length of the observation period (the first canonical variate in the multivariate generalizability analysis) while taking into account the correlations among observation intervals.

Since most observational measurement is linked to some degree, a full treatment of this topic in G theory is clearly needed. Until the theory and procedures for handling linked conditions of a facet are developed, the decisionmaker should, at least, make sure that the time samples in the D study match in duration those in the G study. (For detailed recommendations, see Mitchell, 1979.) Where

observation periods in the G study differ in length or separation in time from those used in the D study, estimates of variance components and generalizability coefficients are likely to be overestimated or underestimated according to some complex function of the magnitude and direction of the correlations among measures from linked observation periods.

In this section we have assumed that the phenomenon being studied remains constant over observations. If this assumption holds, then the linkage among observations is due to correlated error. The problem is much more complex, however, when the universe score changes over time, as is the case in maturation studies (e.g., Bayley, 1968).

This problem is too large to be reviewed here. Among those investigating time-dependent phenomena are Bryk, Strenio, and Weisberg (1980). Although they have not investigated reliability explicitly, they reviewed traditional analysis strategies used in the face of non-equivalent growth systems and suggested alternative methods of analysis.

#### Miscellaneous Topics

Here we mention briefly two other topics. The first topic is signal-to-noise ratios and the second is the relationship between G theory and validity theory.

Signal/Noise Ratios. A signal/noise ratio is defined as the ratio of the universe score variance (signal) to the error variance (noise). It has been proposed by Kane and Brennan (1980) and Tavecchio (1977; Elffers & Tavecchio, 1979) as a means for evaluating the adequacy of a measurement procedure. Brennan and Kane (1977c) discussed this ratio for absolute decisions while Elffers and Tavecchio (1979) discussed it for relative decisions. We mention this topic in passing because we prefer to de-emphasize summary coefficients and emphasize interpretation of the components of error variance in evaluating a measurement procedure.

Relationship between G Theory and Validity Theory. While this topic has been addressed briefly by Cronbach et al. (1972), Cardinet et al. (in press), Fyans (1977), Guttman and Guttman (1976), McDonald (1978) and Van der Kamp (1976), a systematic attempt to integrate G

theory with validity theory has only recently been reported by Kane (1980). Kane's treatment is provocative and elaborate, but too tentative to be covered in this review. We believe, however, that his formulation will set the stage for theoretical developments in the 1980s.

### 3. ILLUSTRATIVE APPLICATION OF GENERALIZABILITY THEORY

In this section, ratings of the educational requirements of occupations are used to illustrate an application of generalizability theory. The study (Webb & Shavelson, 1981; Webb, Shavelson, Shea & Morello, 1981) was conducted by the authors in conjunction with the State of California Employment Department. Specifically, the illustrations include a univariate generalizability analysis, a multivariate generalizability analysis, estimation of variance components with an unbalanced design, and Bayesian estimation of variance components.

#### 3.1. The Study of General Educational Development Ratings

The U.S. Department of Labor developed the General Educational Development (GED) scale to rate the amount of reasoning, mathematics, and language abilities needed to perform various jobs. GED ratings are used in several employment and training situations. For example, they provide the basis for: (a) estimates of time required to learn job skills, (b) state employment agencies' decisions to refer persons to specific employers, job training programs, or remedial education programs, and (c) equating jobs that have similar educational requirements.

In this study, job analysts were given written descriptions of jobs, published in the Dictionary of Occupational Titles, and were asked to rate the jobs on three components of the GED scale: reasoning development, mathematics development, and language development. Each component was measured on a six-point scale. Each of 71 raters from 11 geographic field centers across the U.S. evaluated the three components of a sample of jobs on two occasions. Different centers had different numbers of job analysts, ranging from two to twelve. Hence, the G study design was a partially nested, unbalanced design with different numbers of raters nested within centers. In order to illustrate G theory in its basic form, a random sample of two raters from each center was taken to form a balanced generalizability design.



The design of the study, therefore, was raters (r) nested within geographic centers (c), crossed with jobs (j) and occasions (o). Because we are concerned with estimating the general educational development required to perform each job, the variance component for jobs ( $\hat{\sigma}_j^2$ ) is interpreted as the universe score variance. All other variance components are considered measurement error in this study since absolute decisions are made regarding the GED requirements of each job. These include the components for raters nested within center, center, occasion, and all interactions.

### 3.2 Univariate Generalizability Analysis

For the univariate generalizability study, a random effects analysis of variance was used to estimate the variance components contributing to the observed variation in job ratings. A separate analysis was performed for each component of the GED scale. As recommended by Cronbach et al. (1972), all negative estimates of variance components were replaced with zero in calculating the variance components. For each analysis, the components of variance, the sum of components constituting error variation, and the coefficient of generalizability were computed.<sup>3</sup>

Since this analysis focuses on absolute decisions, the error variance,  $\sigma_A^2$ , reflects not only disagreements about the ordering of the jobs, but also reflects differences in mean ratings. It is important to know, for example, whether raters use essentially the same mean level of the rating scale as well as whether they rank-order jobs similarly.

Data bearing on the generalizability of the ratings of the jobs over occasions, raters, and centers are reported in Table 5 for each of the three GED ratings. The estimated variance components for jobs differ across GED ratings. They suggest that jobs can be distinguished more on their demands for language than on their demands for mathematics and reasoning. The patterns of variance components contributing to error were consistent: raters' ratings accounted for most of the error variation and occasions and centers accounted for little. The patterns of variance components suggest that, by taking the average rating of

Table 4  
Univariate Generalizability Study of G.E.D. Ratings<sup>a</sup>

Source of Variation	Estimated Variance Component		Estimated Variance	
	Reasoning	Mathematics Language	Component with $n_r = 4, n_o = 1, n_c = 1$	Reasoning Mathematics Language
Jobs (J)	.74	.63	1.01	.74 .63 1.01
Occasions (O)	.00	.00	.00	.00 .00 .00
Centers (C)	.00	.02	.05	.00 .02 .05
Raters (Centers) (R:C)	.06	.02	.00	.02 .00 .00
JO	.00	.01	.01	.00 .01 .01
JC	.00	.00	.00	.00 .00 .00
JR:C	.13	.16	.14	.03 .04 .04
OC	.01	.00	.00	.01 .00 .00
OR:C	.00	.09	.07	.00 .02 .02
JOC	.00	.02	.01	.00 .02 .01
JOR:C	.22	.25	.22	.06 .06 .05
$\hat{\sigma}_A^2$	.42	.57	.50	.12 .17 .18
$\hat{\rho}^2$	.64	.53	.67	.86 .79 .85

<sup>a</sup>The design is raters nested within centers crossed with jobs and occasions.

four raters, measurement error can be reduced by about 75% ( $\hat{\sigma}_A^2$  in Table 4) and the generalizability coefficients ( $\hat{\rho}^2$ ) correspondingly increased to .86 for reasoning, .79 for mathematics, and .85 for language.

The consistent patterns of results for the reasoning, mathematics, and language ratings were not unexpected since their correlations are:  $r_{\text{reasoning, math}} = .74$ ;  $r_{\text{reasoning, language}} = .84$ ;  $r_{\text{math, language}} = .73$ . The size of the correlations suggests that all three GED ratings share a common, underlying factor and that a multivariate generalizability coefficient would be appropriate.

### 3.3. Multivariate Generalizability Analysis

For the multivariate generalizability study, a random effects multivariate analysis of variance was performed using the reasoning, mathematics, and language ratings as a vector of scores. Due to the limited capacity of computer programs available to perform the multivariate analysis and because geographic center contributed little to variability among job ratings, geographic center was excluded from the multivariate analysis. The design for this analysis was, then, raters crossed with jobs and occasions.

For each source of variation in the design, variance component matrices were computed from the mean square matrices. Hence, one matrix, for example, comprised estimated universe-score variances and covariances. All matrices with negative estimated variance components (diagonal values) were set equal to zero in further estimation. For this analysis, the matrices of variance components, coefficients of generalizability, and canonical weights corresponding to each coefficient of generalizability were computed.

The estimated variance and covariance component matrices representing the seven sources of variation are presented in Table 5. Only the components for one rater and one occasion are included. To obtain the results for four raters, the components corresponding to the rater main effect and interactions need only to be divided by four.

As a consequence of the calculation procedure, the variance components in Table 5 are the same as those produced by the univariate analysis. The components of covariance, however, provide new information.

Table 5

Estimated Variance and Covariance Components for Multivariate Generalizability  
Study of G.E.D. Ratings ( $n_r = 1$ ,  $n_o = 1$ )<sup>a</sup>

Source of Variation	Reasoning	Mathematics	Language
Jobs (J)	.75		
	.64	.66	
	.88	.74	1.09
Occasions (O)	.00		
	.00	.00	
	.00	.00	.00
Raters (R)	.03		
	.03	.09	
	.03	.05	.05
JO	.00		
	.00	.00	
	.00	.00	.00
JR	.12		
	.11	.13	
	.09	.07	.11
OR	.00		
	.01	.01	
	.00	.00	.01
JRO,e	.21		
	.07	.29	
	.11	.10	.26

<sup>a</sup>The design is raters crossed with jobs and occasions.

The large components for jobs reflect the underlying correlations among the GED components. Jobs that require high reasoning ability are seen by the raters to require high mathematics and language ability. Whereas the nonzero components of variance for raters indicate that some raters give higher ratings than others, the positive components of covariance indicate that the raters who give higher ratings on one GED component are likely to give higher ratings on the other GED components. The positive components for the job x rater interaction suggest that not only do raters disagree about which jobs require more ability, but their disagreement is consistent across GED components. The nonzero components for error suggest that the unexplained factors that contribute to the variation of ratings also contribute to the covariation between ratings. As expected, the components of covariance due to the occasion main effect and interactions are negligible.

Composites of general educational development that have maximum generalizability are presented in Table 6. When the generalizability of GED ratings was estimated for one rater and one occasion, one dimension with a generalizability coefficient exceeding .50 emerged from the analysis. This dimension is a verbal composite of reasoning and language. The analysis using four raters and one occasion produced two dimensions with generalizability coefficients exceeding .50. The first composite is defined by reasoning and language. This composite has a generalizability coefficient of .74 for one rater and .92 for four raters. As in the univariate case, the estimate of measurement error is reduced by 75% when four raters are used. The second composite is a contrast between mathematics and language or, using more common terminology, a verbal-quantitative contrast. The estimate of generalizability for this contrast is .62 for a D study with four raters and one occasion.

### 3.4. Unbalanced Designs

The original design of the study was unbalanced. This section illustrates the estimation of variance components for an unbalanced design. The unbalanced design analyzed here is raters nested within a random sample of five of the eleven geographic centers crossed with

Table 6

CANONICAL VARIATES FOR MULTIVARIATE GENERALIZABILITY  
STUDY OF G.E.D. RATINGS<sup>a</sup>

G.E.D. Component	Canonical Coefficients		
	$n_r = 1, n_o = 1$	$n_r = 4, n_o = 1$	
	I	I	II
Reasoning	.34	.38	.05
Mathematics	.06	.06	-1.95
Language	.51	.57	1.33
Coefficient of Generalizability ( $\rho^2$ )	.74	.92	.62

<sup>a</sup>The design is raters crossed with jobs and occasions.

jobs and occasions. (The restriction to five centers will be explained below.) The results of this analysis will be compared to those of two balanced designs: (1) two raters randomly sampled from each of the five centers, and (2) two raters randomly selected from each of the 11 centers.

The estimates of variance components in the unbalanced design were obtained using a modification of Rao's MIVQUE procedure suggested by Hartley, Rao, and LaMotte (1978; see section 2.1.5). Because the SAS procedure VARCOMP would require an excessive amount of region if all 11 geographic centers were to be included in the analysis, only a subset of the centers could be used. Limiting the amount of region required to perform the analysis to 300K bytes of core, the largest design that the computer program would run had five centers with a total of 28 raters. For these five centers, the number of raters per center ranged from two to seven. The results of the three analyses--unbalanced design with five centers, balanced design with five centers, and balanced design with 11 centers--are presented in Table 8. The estimates of the variance components are much the same in the three analyses (see Table 8). The primary source of variability, as was seen in the applications

Table 7

Estimates of Variance Components from Unbalanced and Balanced Data  
for G.E.D. Ratings of Reasoning Ability<sup>a</sup>

Source of Variation	5 centers		11 centers	
	Unbalanced Design (2 to 7 raters per center)	Balanced Design (2 raters per center)	Balanced Design (2 raters per center)	
Jobs (J)	.74	.67	.74	
Occasions (o)	.00	.00	.00	
Centers (c)	.00*	.00*	.00	
Raters (Centers)(R:C)	.07	.07	.06	
JO	.00	.00	.00	
JC	.00	.00	.00	
JR:C	.11	.08	.13	
OC	.00*	.01	.01	
OR:C	.02	.00*	.00	
JOC	.00*	.00*	.00*	
JOR:C	.20	.24	.22	
$\hat{\sigma}_A^2$	.40	.40	.42	
$\hat{\rho}^2$	.65	.63	.64	

<sup>a</sup>The design is raters nested within centers crossed with jobs and occasions.

discussed previously, was raters' ratings. The estimates of error variance ( $\hat{\sigma}_\Delta^2$ ) and generalizability coefficients ( $\hat{\rho}^2$ ) are also similar across the three designs.

A few minor differences are apparent across the three analyses. First, the variance component for jobs was smaller in the balanced design with five centers than in the other two designs. Apparently, the raters in this design used a smaller range in their ratings of jobs than did the raters analyzed in the other designs. This discrepancy did not, however, change the overall results concerning the dependability of raters' judgments. The second difference across analyses was the negative variance components. The analyses of five centers produced more negative estimates of variance components than the analysis of the 11 centers. The patterns of negative estimates, however, were similar across the analyses. Components involving geographic center were the most likely candidates for negative estimation.

To determine whether the five centers analyzed in the unbalanced design might have produced atypical results compared to analyses using other subsets of centers, five additional analyses were carried out using other combinations of centers. The analyses were subject to the limit of 300K bytes of core. All of the additional analyses produced nearly the same results as those reported in Table 7. The patterns of variance components and the resulting estimates of error variance and generalizability coefficients were very similar. Across five additional analyses, the coefficients of generalizability ranged from .61 to .66.

With the present methodology, two strategies seem to be available for analyzing reasonably large unbalanced designs: (1) to sample conditions of the nested facet to produce a balanced (crossed) design, or (2) to reduce the unbalanced design. In the analyses discussed here, the two approaches yielded similar estimates of the components of variance. Although the two options may produce similar results, the first option, sampling to produce a crossed design, affords greater flexibility in choosing computational procedures.

### 3.5. Bayesian Estimation

The Bayesian estimates of modal variance components, presented in Section 2.1.2, assume a non-informative prior which includes the constraint



that estimated variance components cannot be zero. Using [16], modal estimates can readily be calculated from the sums of squares provided by an ANOVA. In general, the adjusted mean square is given by:

$$\hat{V} = \frac{SS}{df+2}$$

The Bayesian modal estimates of the variance components ( $\hat{\sigma}^2$ ) can be obtained by equating the  $\hat{V}$ s--the adjusted mean squares--to their expectations. For example, consider the Job x Occasion x Center x Rater: Center generalizability study and the GED rating for reasoning (see Table 4). The sums of squares for the residual (JOC:C) was 63.01 and  $df_{res}$  was 286. The adjusted mean square is:  $\hat{V}_{res} = 63.01/(286 + 2) = .219$  and so  $\hat{\sigma}_{res}^2 = .219$ , while  $\hat{\sigma}_{res}^2 = .220$  (Table 8). The sums of squares for the next source of variation, JOC, was 48.501 and  $df_{JOC}$  was 260. The adjusted mean square is:  $\hat{V}_{JOC} = 48.501/(260 + 2) = .185$ . Setting  $\hat{V}_{JOC}$  equal to its expectation, we obtain  $\hat{\sigma}_{JOC}^2 = -.0175$ .

Since Bayesian estimates are constrained to be greater than or equal to zero, the negative value of  $\hat{\sigma}_{JOC}^2$  indicates that  $\hat{V}_{JOC}$  provides a second estimate of error independent of the estimate provided by  $\hat{\sigma}_{res}^2$ . That is, the expected value of  $\hat{V}_{JOC}$  is  $\hat{\sigma}_{res}^2 + n_r \hat{\sigma}_{JOC}^2$ . Since  $\hat{\sigma}_{JOC}^2$  is constrained to be zero and not negative,  $\hat{V}_{JOC} = \hat{\sigma}_{res}^2$ . The Bayesian approach then pools the two estimates of measurement error as follows:

$$\hat{V}_{res(pooled)} = \frac{SS_{res} + SS_{JOC}}{df_{res} + df_{JOC}} .$$

Setting  $\hat{V}_{res}$  equal to its expectation,  $\hat{\sigma}_{res(pooled)}^2 = .203$ . This pooled estimate is carried through subsequent calculations of variance components and the generalizability coefficient. It is also used in interpreting the results of the G study.

In Table 8, the Bayesian estimates of the modal variance components are compared with the traditional estimates. As expected, the Bayesian estimates are slightly smaller than the traditional estimates. The

Table 8  
Comparison of Bayesian and Traditional  
Estimates of Variance Components: Ratings  
of Reasoning<sup>a</sup>

Source	Bayesian Estimates of Variance Components	Traditional Estimates of Variance Components
Jobs (J)	.69	.74
Occasions(O)	.00	.00
Centers(C)	.00	.00
Raters(Centers)(R:C)	.05	.06
JO	.00	.00
JC	.00	.00
JR:C	.13	.13
OC	.01	.01
OR:C	.00	.00
JOC	.00	.00
JOR:C(res)	.20	.22
Error Variance ( $\hat{\sigma}_\Delta^2$ )	.39	.42
Generalizability ( $\hat{\rho}^2$ )	.64	.64

<sup>a</sup>The design is raters nested within centers crossed with jobs and occasions.

Bayesian generalizability coefficient, however, is equal to the traditional estimate.

While the two procedures for calculating the estimates differ little, the assumptions underlying the two estimation approaches differ considerably. Perhaps the major difference, in addition to the non-informative prior for the Bayesian estimates, is the pooling procedure associated with the Bayesian estimates. This procedure makes use of the information available when a negative estimate arises, something the traditional theory, in practice, ignores (see Box & Tiao, 1973, on problems of pooling with the traditional approach).

# REFERENCES

- Bayley, N., "Behavioral Correlates of Mental Growth: Birth to Thirty-six Years," American Psychologist, 1968, 23, 1-17.
- Bock, R. D., "Multivariate Analysis of Variance of Repeated Measurements," in C. W. Harris (Ed.), Problems of Measuring Change, Madison, Wisconsin: University of Wisconsin Press, 1963.
- , "Contributions of Multivariate Experimental Designs to Educational Research," in R. B. Cattell (Ed.), Handbook of Multivariate Experimental Psychology, Chicago: Rand McNally, 1966.
- , Multivariate Statistical Methods in Behavioral Research, New York: McGraw-Hill, Inc., 1975.
- Box, G. E. P., and G. C. Tiao, Bayesian Inference in Statistical Analysis, Reading, Mass.: Addison-Wesley, 1973.
- Brennan, R. L., "The Calculation of Reliability from a Split-plot Factorial Design," Educational and Psychological Measurement, 1975, 35, 779-788.
- , Generalizability Analyses: Principles and Procedures (ACT Technical Bulletin No. 26), Iowa City, Iowa: American College Testing Program, September 1977 (a).
- , KR-21 and Lower Limits of an Index of Dependability for Mastery Tests (ACT Technical Bulletin No. 27), Iowa City, Iowa: American College Testing Program, December 1977 (b).
- , Algorithms, Procedures, and Variance Components for Analysis of Variance (ACT Technical Bulletin No. 31), Iowa City, Iowa: American College Testing Program, June 1978 (a).
- , Extensions of Generalizability Theory to Domain-referenced Testing (ACT Technical Bulletin No. 30), Iowa City, Iowa: American College Testing Program, June 1978 (b).
- , Handbook for Gapid: A Fortran IV Computer Program for Generalizability Analyses with Single Facet Designs (ACT Technical Report No. 34), Iowa City, Iowa: American College Testing Program, October 1979 (a).
- , Some Applications of Generalizability Theory of the Dependability of Domain-referenced Tests (ACT Technical Bulletin No. 32), Iowa City, Iowa: American College Testing Program, April 1979 (b).
- , "Applications of Generalizability Theory," in R. A. Berk (Ed.), Criterion-referenced Measurement: State of the Art, Baltimore: Johns Hopkins Press, 1980.

- Brennan, R. L., D. Jarjoura, and E. L. Deaton, Interpreting and Estimating Variance Components in Generalizability Theory: An Overview, paper presented at the annual meeting of the American Educational Research Association, Boston, April 1980.
- Brennan, R. L., and M. T. Kane, "An Index of Dependability for Mastery Tests," Journal of Educational Measurement, 1977, 14, 277-289 (a).
- , "Signal/noise Ratios for Domain-referenced Tests, Psychometrika, 1977, 42, 609-625 (b).
- , "Generalizability Theory: A Review of Basic Concepts, Issues and Procedures," in R. E. Traub (Ed.), New Directions in Testing and Measurement, San Francisco: Jossey-Bass, 1979.
- Brennan, R. L., and R. E. Lockwood, A Comparison of Two Cutting Score Procedures Using Generalizability Theory, (ACT Technical Bulletin No. 33), Iowa City, Iowa: American College Testing Program, April 1979.
- Byrk, A. S., J. F. Strenio, and H. I. Weisberg, "A Method for Estimating Treatment Effects when Individuals are Growing," Journal of Educational Statistics, 1980, 5, 5-34.
- Calkins, D. S., O. Erlich, P. T. Marston, and D. Malitz, An Empirical Investigation of the Distributions of Generalizability Coefficients and Various Estimates for an Application of Generalizability Theory, paper presented at the annual meeting of the American Educational Research Association, Toronto, March 1978.
- Cardinet, J., and Y. Tourneur, "The Facets of Differentiation [sic] and Generalization in Test Theory" (shortened English version of the original text: "Une theorie des tests pedagogiques"), paper presented at the 18th Congress of the International Association of Applied Psychology, Montreal, July-August 1974.
- , "How to Structure and Measure Educational Objectives in Periodic Surveys," in R. Summer (Ed.), Monitoring National Standards of Attainment in Schools, National Foundation for Educational Research in England and Wales, Slough, England, June 1977.
- , "Le Calcul de Marges d'Erreurs Dans la Theorie de la Generalizabilite," Neuchatel (Suisse): Institut Romand de Recherches et de Documentation Pedagogiques, 1978.
- Cardinet, J., Y. Tourneur, and L. Allal, "The Generalizability of Surveys of Educational Outcomes," in D. N. M. De Gruijter, and L. J. Th. van der Kamp (Eds.), Advances in Psychological and Educational Measurement, New York: Wiley, 1976, 185-198 (a).
- , "The Symmetry of Generalizability Theory: Applications to Educational Measurement," Journal of Educational Measurement, 1976, 13, 119-135 (b).

- Cardinet, J., Y. Tourneur, and L. Allal, "Enlargement of Generalizability Theory and Its Applications in Educational Measurement," Journal of Educational Measurement, in press.
- Coates, T. J., M. R. Rosekind, R. J. Strossen, C. E. Thoresen, and K. Kirmil-Gray, "Sleep Recordings in the Laboratory and Home: A Comparative Analysis," Psychophysiology, 1979, 16, 339-347.
- Conger, A. J., "Estimating Profile Reliability and Maximally Reliable Composites," Multivariate Behavioral Research, 1974, 9, 85-104.
- Conger, A. J., and R. Lipshitz, "Measures of Reliability of Profiles and Test Batteries," Psychometrika, 1973, 38, 411-427.
- Cornelius, E. T., J. A. Woodward, and R. G. Demaree, CRONB: A Fortran IV Program to Compute Variance Components for Various Experimental Designs, Texas Christian University Institute of Behavioral Sciences, 1976.
- Cornfield, J., and J. W. Tukey, "Average Values of Mean Squares in Factorials," Annals of Mathematical Statistics, 1956, 27, 907-949.
- Cronbach, L. J., "On the Design of Educational Measures," in D.N.M. De Gruijter and L. J. Th. van der Kamp (Eds.), Advances in Psychological and Educational Measurement, New York: Wiley, 1976.
- Cronbach, L. J., and P. J. D. Drenth, (Eds.), Mental Tests and Cultural Adaptation, The Hague, Netherlands: Mouton, Inc., 1972.
- Cronbach, L. J., and L. Furby, "How Should We Measure Change--or Should We?," Psychological Bulletin, 1970, 74, 68-80.
- Cronbach, L. J., G. C. Gleser, H. Nanda, and N. Rajaratnam, The Dependability of Behavioral Measurements, New York: Wiley, 1972.
- Cronbach, L. J., N. Rajaratnam, and B. Gleser, "Theory of Generalizability: A Liberalization of Reliability Theory," British Journal of Statistical Psychology, 1963, 16, 137-163.
- Darlington, R. B., "Multiple Regression in Psychological Research and Practice," Psychological Bulletin, 1978, 3, 161-182.
- Davis, C., Bayesian Inference in Two Way Models: An Approach to Generalizability, unpublished doctoral dissertation, University of Iowa, 1974.
- de Finetti, B., "Foresight: Its Logical Laws, Its Subjective Sources," in H. E. Kyburg, and G. E. Smokler (Eds.), Studies in Subjective Probability, New York: Wiley, 1964.
- Edinberg, M. A., P. Karoly, and G. C. Gleser, "Assessing Assertion in the Elderly: An Application of the Behavioral-analytic Model of Competence," Journal of Clinical Psychology, 1977, 33, 869-874.

- Elffers, H., and L. W. C. Tavecchio, Variance Components in Test Generalizability Research: Which, When, Why?, Utrecht/Leiden: Vereniging voor On Derwijs Research (Dutch Association for Educational Research), VOR-Publikatie 9, April 1979.
- Erlich, O., and G. Borich, "Factorial Analysis of Generalizability," Educational and Psychological Measurement, 1978, 38, 125-133.
- Erlich, O., and R. Shavelson, "Generalizability of Measures: A Computer Program for Two and Three Facet Designs," Behavior Research Methods and Instrumentation, 1976, 8, 275 (a).
- , The Application of Generalizability Theory to the Study of Teaching (Technical Report 76-9-1), "Beginning Teacher Evaluation Study," Far West Laboratory, September 1976 (b).
- , "The Search for Correlations between Measures of Teacher Behavior and Student Achievement: Measurement Problem, Conceptualization Problem or Both?" Journal of Educational Measurement, 1978, 15, 77-89.
- Farrell, A. D., J. J. Marco, A. J. Conger, J. P. Curran, and J. L. Wallander, "Self-ratings and Judges Ratings of Heterosexual Social Anxiety: A Generalizability Study," Journal of Consulting and Clinical Psychology, 1979, 47, 164-175.
- Feller, W., An Introduction to Probability Theory and Its Applications, New York: Wiley, 1966.
- Fyans, L. J., Jr., A New Multiple Level Approach to Cross-cultural Psychological Research, unpublished doctoral dissertation, University of Illinois at Urbana-Champaign, 1977.
- Fyans, L. J., F. Salili, M. L. Maehr, and K. A. Desai, Cultural variation in the Meaning of Achievement, paper presented at the annual meeting of the American Educational Research Association, Boston, April 1980.
- Gershen, J. A., An Evaluation of the Small Group Instructional Methods for Teaching Behavioral Sciences in the Dental Curriculum, unpublished doctoral dissertation, University of California, Los Angeles, 1976.
- Gillmore, G. M., An Introduction to Generalizability Theory as a Contributor to Evaluation Research, Seattle, Wash.: Educational Assessment Center, University of Washington, March 1979.
- , Student Instructional Ratings: To What Universe Can We Dependably Generalize Results?, paper presented at the annual meeting of the American Educational Research Association, Boston, April 1980.

- Gillmore, G. M., M. T. Kane, and R. W. Naccarato, "The Generalizability of Student Ratings of Instruction," Journal of Educational Measurement, 1978, 15, 1-15.
- Glaser, R., "Instructional Technology and the Measurement of Learning Outcomes," American Psychologist, 1963, 18, 519-521.
- Gleser, G. C., B. L. Green, and C. N. Winget, "Quantifying Interview Data on Psychic Impairment of Disaster Survivors," The Journal of Nervous and Mental Diseases, 1978, 166, 209-216.
- Guttman, L., and R. Guttman, "A Theory of Behavioral Generality and Specificity During Mild Stress," Behavioral Science, 1976, 21, 469-477.
- Haggard, E. A., Interclass Correlation and the Analysis of Variance, New York: Dryden, 1958.
- Hambleton, R. K., and M. R. Novick, "Toward an Integration of Theory and Method of Criterion-referenced Tests," Journal of Educational Measurement, 1973, 10, 159-170.
- Hambleton, R. K., H. Swaminathan, J. Algina, and B. C. Courson, "Criterion-referenced Testing and Measurement: A Review of Technical Issues and Developments," Review of Educational Research, 1978, 48, 1-47.
- Harnqvist, K., "Canonical Analyses of Mental Test Profiles," Scandinavian Journal of Psychology, 1973, 14, 282-290.
- Harris, C. W., "An Interpretation of Livingston's Reliability Coefficient for Criterion-referenced Tests," Journal of Educational Measurement, 1972, 9, 27-29.
- Hartley, H. O., J. N. K. Rao, and L. LaMotte, "A Simple Synthesis-based Method of Variance Component Estimation," Biometrics, 1978, 34, 233-242.
- Henderson, C. R., "Estimation of Variance and Covariance Components," Biometrics, 1963, 9, 226-252.
- Hill, B. M., "Some Contrasts Between Bayesian and Classical Influence in the Analysis of Variance and the Testing of Models," in D. L. Meyer and R. O. Collier, Jr. (Eds.), Bayesian Statistics, Ithaca, Ill.: F. E. Peacock, 1970.
- , "Correlated Errors in the Random Model," Journal of the American Statistical Association, 1967, 62, 1387-1400.
- , "Inference about Variance Components in the One-way Model," Journal of the American Statistical Association, 1965, 60, 806-825.



Huysamen, G. K., Psychological Test Theory, Duranville, South Africa: Uitgewery Boschendal Distributor, 1980.

Joe, G. N., and J. A. Woodward, "Some Developments in Multivariate Generalizability," Psychometrika, 1976, 41, 205-217.

Joreskog, K. G., "A General Approach to Confirmatory Maximum Likelihood Factor Analysis," Psychometrika, 1969, 34, 183-201.

-----, "Statistical Analysis of Sets of Cogeneric Tests," Psychometrika, 1971, 36, 109-133.

-----, "Analyzing Psychological Data by Structural Analysis of Covariance Matrices," in D. H. Krantz, R. C. Atkinson, R. D. Luce, and P. Suppes (Eds.), Contemporary Developments in Mathematical Psychology (Vol. II), W. H. Freeman & Company, 1974.

Kane, M. T., Interpreting Variance Components as Evidence for Reliability and Validity, paper presented at the annual meeting of the American Educational Research Association, Boston, April 1980.

Kane, M. T., and R. L. Brennan, "The Generalizability of Class Means," Review of Educational Research, 1977, 47, 267-292.

-----, "Agreement Coefficients as Indices of Dependability for Domain-referenced Tests," Applied Psychological Measurement, 1980, 4, 105-126.

Kane, M. T., G. M. Gillmore, and T. J. Crooks, "Student Evaluations of Teaching: The Generalizability of Class Means," Journal of Educational Measurement, 1976, 13, 171-183.

Katerberg, R., F. J. Smith, and S. Hoy, "Language Time and Person Effects on Attitude Scale Translations," Journal of Applied Psychology, 1977, 62, 385-391.

Kelley, T. L., Fundamentals of Statistics, Cambridge, Mass.: Harvard University Press, 1947.

Kingman, J. F. C., "Uses of Exchangeability," The Annals of Probability, 1978, 6, 183-197.

Leone, F. C., and L. S. Nelson, "Sampling Distributions of Variance Components--I. Empirical Studies of Balanced Nested Designs," Technometrics, 1966, 8, 457-468.

Lewis, C., M. Wang, and M. R. Novick, "Marginal Distributions for the Estimation of Proportions in m Groups," Psychometrika, 1975, 40, 63-75.

Lindley, D. V., "The Estimation of Many Parameters," in V. P. Godambe and D. A. Sprott (Eds.), Foundations of Statistical Inference, Toronto, Canada: Holt, Rinehart & Winston, 1971.

Lindley, D. V., and M. R. Novick, The Role of Exchangeability in Inference (Technical Report No. 79-8), University of Iowa, Division of Educational Psychology, August 1979.

Lindquist, E. F., Design and Analysis of Experiments in Psychology and Education, Boston: Houghton-Mifflin, 1953.

Linn, R. L., "Issues of Reliability in Measurement for Competency-based Programs," in M. A. Bunda and J. R. Sanders (Eds.), Practices and Problems in Competency-based Educations, Washington, D.C.: NCME Monograph, 1979.

Livingston, S. A., "Criterion-referenced Applications of Classical Test Theory," Journal of Educational Measurement, 1972, 9, 13-26 (a).

-----, "Reply to Shavelson, Block and Ravitch's 'Criterion-referenced Testing: Comments on Reliability,'" Journal of Educational Measurement, 1972, 9, 139-140(b).

Loevinger, J., "Person and Population as Psychometric Concepts," Psychological Review, 1965, 72, 143-155.

Llabre, M. M., An Application of Generalizability Theory to the Assessment of Writing Ability, unpublished doctoral dissertation, University of Florida, 1978.

-----, Estimating Variance Components with Unbalanced Designs in Generalizability Theory, paper presented at the annual meeting of the American Educational Research Association, Boston, April 1980.

McDonald, R. P., "Generalizability in Factorable Domains: Domain Validity and Generalizability," Educational and Psychological Measurement, 1978, 38, 75-79.

Malgady, R. G., J. A. Amato, and S. W. Huck, "The Fixed-effect Fallacy in Educational Psychological Research: A Problem of Generalizing to Multiple Populations," Educational Psychologist, 1979, 14, 79-86.

Mariotto, M. J., and A. D. Farrell, "Comparability of the Absolute Level of Ratings on the Inpatient Multidimensional Psychiatric Scale Within a Homogeneous Group of Raters," Journal of Consulting and Clinical Psychology, 1979, 47, 59-64.

Mitchell, S. K., "Interobserver Agreement, Reliability, and Generalizability of Data Collected in Observational Studies," Psychological Bulletin, 1969, 86, 376-390.

Molenaar, I. W., and C. Lewis, An Improved Model and Computer Program for Bayesian m-group Regression (Technical Report 79-5), Division of Educational Psychology, University of Iowa, March 1979.

Novick, M. R., "Multiparameter Bayesian Inference Procedures" (with discussion), Journal of the Royal Statistical Society: Series B, 1969, 31, 29-64.

-----, "Bayesian Methods in Educational Testing: A Third Survey," in D. M. N. de Gruijter and L. J. Th. van der Kamp (Eds.), Advances in Psychological and Educational Testing, New York, Wiley, 1976.

Novick, M. R., and W. J. Hall, "A Bayesian Indifference Procedure," Journal of the American Statistical Association, 1965, 60, 1104-1117.

Novick, M. R., and P. H. Jackson, "Bayesian Guidance Technology," Review of Educational Research, 1970, 40, 459-494.

-----, Statistical Methods for Educational and Psychological Research, New York: McGraw-Hill, 1974.

Novick, M. R., P. H. Jackson, and D. T. Thayer, "Bayesian Inference and the Classical Test Theory Model: Reliability and True Scores," Psychometrika, 1971, 36, 261-288.

Novick, M. R., P. H. Jackson, D. T. Thayer, and N. S. Cole, "Estimating Multiple Regressions in m Groups: A Cross Validation Study," British Journal of Mathematical and Statistical Psychology, 1972, 25, 33-50.

Novick, M. R., C. Lewis, and P. H. Jackson, "The Estimation of Proportions in m Groups," Psychometrika, 1973, 38, 19-46.

Peng, S. S., and S. D. Farr, "Generalizability of Free-recall Measurements," Multivariate Behavioral Research, 1976, 11, 287-296.

Popham, W. J., Educational Evaluation, Englewood Cliffs, N. J.: Prentice-Hall, 1975.

Rao, C. R., "Minimum Variance Quadratic Unbiased Estimation of Variance Components," Journal of Multivariate Analysis, 1971, 1, 445-456.

-----, "Estimation of Variance and Covariance Components in Linear Models," Journal of the American Statistical Association, 1972, 67, 112-115.

Rowley, G., "The Reliability of Observational Measures," American Educational Research Journal, 1976, 13, 51-59.

-----, "The Relationship of Reliability in Classroom Research to the Amount of Observation: An Extension of the Spearman-Brown Formula," Journal of Educational Measurement, 1978, 15, 165-180.

"SAS Institute," SAS User's Guide, Raleigh, N. C.: SAS Institute, Inc., 1979.

- Scheffe, H., The Analysis of Variance, New York: Wiley, 1959.
- Searle, S. R., Linear Models, New York: Wiley, 1971.
- Shavelson, R. J., "A Criterion Sampling Approach to Measurement in the Evaluation of Nonformal Education," UNESCO Institute for Education, Hamburg, Germany, December 1979.
- Shavelson, R., J. Block, and M. Ravitch, "Criterion-referenced Testing: Comments on Reliability," Journal of Educational Measurement, 1972, 9, 133-137.
- Smith, P., "Sampling Errors of Variance Components in Small Sample Multifacet Generalizability Studies," Journal of Educational Statistics, 1978, 3, 319-346.
- Smith, P. L., "The Generalizability of Student Ratings of Courses: Asking the Right Questions," Journal of Educational Measurement, 1979, 16, 77-88.
- , Some Approaches to Determining the Stability of Estimated Variance Components, paper presented at the annual meeting of the American Educational Research Association, Boston, April 1980.
- Tavecchio, L. W. C., "Quantification of Teaching Behavior in Physical Education: A Methodological Study," (Doctoral dissertation, University of Amsterdam, 1977), (University Microfilms International No. 77-70, 039).
- Tourneur, Y., Les Objectifs du Domaine Cognitif, 2me Partie--Theorie des Tests, Ministere de l'Education Nationale et de la Culture Francaise, Universite de l'Etat a Mons, Faculte des Sciences Psycho-Pedagogiques, 1978.
- Tourneur, Y., and J. Cardinet, Analyse de Variance et Theorie de la Generalizabilite: Guide Pour la Realisation des Calculs (Document 790.803/CT/9), Universite de l'Etat a Mons, 1979.
- Travers, K. J., Correction for Attenuation: A Generalizability Approach Using Components of Covariance, unpublished manuscript, University of Illinois, 1969.
- U. S. Department of Labor, Handbook for Analyzing Jobs, Washington, D. C.: U. S. Government Printing Office, 1972.
- van der Kamp, L. J. Th., "Generalizability and Educational Measurement," in D. N. M. De Gruijter and L. J. Th. van der Kamp (Eds.), Advances in Psychological and Educational Measurement, New York: Wiley, 1976.
- Webb, N. W., Implementation of BTES Interaction Activities in Junior High School Mathematics, final report submitted to the California Commission for Teacher Preparation and Licensing, 1980.

- Webb, N. M., R. J. Shavelson, J. Shea, and E. Morello, "Generalizability of General Educational Development Ratings of Jobs in the U.S.," Journal of Applied Psychology, 1981, 66, 186-192.
- Webb, N. M., and R. J. Shavelson, "Multivariate Generalizability of General Educational Development Ratings," Journal of Educational Measurement, 1981, 18(1), 13-22.
- Wiggins, J. S., Personality and Prediction: Principles of Personality Assessment, Reading, Mass.: Addison-Wesley, 1973.
- Wilcox, R. R., "Estimating the Likelihood of False-positive and False-negative Decisions in Mastery Testing: An Empirical Bayes Approach," Journal of Educational Statistics, 1977, 2, 289-307.
- Wilcox, R. R., "Estimating True Score in the Compound Binomial Error Model," Psychometrika, 1978, 43, 245-258.
- , "On False-positive and False-negative Decisions with a Mastery Test," Journal of Educational Statistics, 1979, 4, 59-73.
- Wood, R., "Halo and Other Effects in Teacher Assessments," Durham Research Review, 1976, 36, 1120-1126 (a).
- , "Trait Measurement and Item Banks," in D. M. N. de Gruijter and L. J. Th. van der Kamp (Eds.), Advances in Psychological and Educational Measurement, New York: Wiley, 1976 (b).
- Woodward, J. A., and G. W. Joe, "Maximizing the Coefficient of Generalizability in Multi-facet Decision Studies," Psychometrika, 1973, 38, 173-181.

FOOTNOTES

The authors wish to acknowledge the extensive comments of Lee J. Cronbach on each of the earlier drafts of this review. In some cases, they led us to re-think major points in the review and they almost always helped us to clarify our thinking. We also wish to acknowledge the comments of Jean Cardinet, Robert Brennan, and Philip Levy on an early draft of the review; they also helped us to clarify our thinking. We also wish to acknowledge the help of Linda Allal, Oded Erlich, Leslie Fyans, Gerald Gillmore, Michael Kane, Glenn Rowley, and Philip Smith. Finally, we are indebted to Cindy Ornest for typing the numerous drafts of the review.

<sup>1</sup>Introductions to G theory are provided by Brennan (1977a, 1979a, Brennan and Kane (1980), Erlich and Shavelson (1976b), Gillmore (1979), Cardinet and Tourneur (1978), Huysamen (1980), Tourneur (1978), Tourneur and Cardinet (1977), Van der Kamp (1976), and Wiggins (1973).

<sup>2</sup>For brief discussions of G theory and its application to criterion-referenced measurement not treated here, see Cardinet and Tourneur (1974), Cardinet, Tourneur and Allal (in press), Cronbach (1976), Davis (1974), Kane and Brennan (1977), and Tourneur (1977).

<sup>3</sup>Computer programs designed specifically for univariate generalizability analyses have been reported by Brennan (1979a), Cornilius, Woodward, and Demaree (1976), Erlich and Shavelson (1976a), and Erlich and Borich (1978).

